"Sampling and Computation of Weights in the MAFE Surveys"

Schoumaker, Bruno ; mezger

Document type : *Document de travail (Working Paper)*

MAFE Migrations between Africa and Europe

MAFE Methodological Note 6

Sampling and Computation Weights in the MAFE Surveys

SCHOUMAKER Bruno (UCL) and MEZGER Cora (INED),

with the collaboration of

RAZAFINDRATSIME Nicolas (INED) and BRINGE Arnaud (INED)

January 2013

ined INSTITUT NATIONAL D'ÉTUDES DÉMOGRAPHIQUES

ANR

îledeFrance

Funded under the Socio-economic Sciences & Humanities Theme

SEVENTH FRAMEWORK PROGRAMME

The MAFE project is the product of a collective effort. It is coordinated by INED (C. Beauchemin) and is formed, additionally by the Université catholique de Louvain (B. Schoumaker), Maastricht University (V. Mazzucato), the Université Cheikh Anta Diop (P. Sakho), the Université de Kinshasa (J. Mangalu), the University of Ghana (P. Quartey), the Universitat Pompeu Fabra (P. Baizan), the Consejo Superior de Investigaciones Científicas (A. González-Ferrer), the Forum Internazionale ed Europeo di Ricerche sull'Immigrazione (E. Castagnone), and the University of Sussex (R. Black).

All people involved in the survey conception and data collection are cited in Appendix 1. The INED survey department provided its expertise for the original design of the survey. The MAFE project received funding from the European Community's Seventh Framework Programme under grant agreement 217206. In addition, the MAFE-Senegal survey was conducted with the financial support of INED, the Agence Nationale de la Recherche (France), the Région Ile de France and the FSP programme 'International Migrations, territorial reorganizations and development of the countries of the South'. For more details, see: http://www.mafeproject.com/

CONTENTS

INTRODUCTION

This note describes the sampling strategy and the computation of weights in the MAFE surveys. The first part explains designs in Ghana, DR Congo and Senegal, the second part describes the computation of weights in the European samples, followed by the description of the computation of normalized weights to be used with pooled data sets in the third part. The fourth part provides a short review of the literature about the use of weights in different types of analysis. The review provides the background to the final section, which contains indications regarding use of weights for analysis with MAFE data.

For MAFE surveys in Africa as well as in Europe, we present first a general description of the sampling design and a brief explanation of the computation of weights. Sampling methods and the computation of weights is then presented in detail for each country and migration flow. As a result, there are some repetitions in the various parts of the documents - this is done purposefully so that readers interested in one specific country or in the overall approach have the essential information in one part of the paper.

Annexes present (1) the selection grid for individuals within households, and (2) the stata syntax used for sampling households within the selected enumeration areas in Ghana, and (3) the stata syntax used to compute weights in Ghana.

BRIEF DESCRIPTION OF SAMPLING AND THE COMPUTATION OF WEIGHTS IN AFRICA

SAMPLING IN AFRICA

This section presents the general sampling methodology for the household and biographic MAFE surveys conducted in Ghana (2009), DR Congo (2009), and Senegal (2008). The main elements are summarized in Table 1.

In all three countries, stratified multi-stage random samples of households and individuals in the target areas were selected. The target areas were the city of Kinshasa in DR Congo, the region of Dakar in Senegal and two cities (Accra and Kumasi) in Ghana. In each of these cities, a sampling frame of primary sampling units was prepared.

In Ghana and Senegal, relatively recent censuses (2000 and 2002 respectively) were available and served as sampling frames at the first stage. In DR Congo, no recent census was available (the latest census was conducted in 1984). Therefore, the sampling frame of the 2007 DHS was used to select a sample of 29 neighbourhoods (out of 324) with a probability proportional to size, and 3 streets were selected randomly with a probability proportional to size in each selected neighbourhood. The sample was stratified at the first stage. Three strata were distinguished in Kinshasa, based on the prevalence of migration. Two strata were considered in Ghana, corresponding to the two cities (Accra and Kumasi). 80 census enumeration areas were randomly selected in Ghana, with a probability proportional to size (60 in Accra, and 20 in Kumasi). In Senegal, the 2 109 c ensus districts in the region of Dakar were divided into 10 strata of equal size based on the migration prevalence (number of households with at least one migrant according to the definition used by the census). In each of the 10 strata, 6 census districts were sampled with a probability proportional to size in terms of the number of households in the district to give a sample of 60 census districts.

In all three countries, a listing operation was carried out in each of the selected survey sites (enumeration areas/census districts or streets) to prepare the sampling frame of households. The listing consisted in enumerating all the households in the selected sites, and in identifying whether these households included migrants or not. In Ghana and DR Congo, three strata were constituted at this stage: households with return migrants, households with migrants abroad, and households without migrants. In Senegal, the process was similar, but only two strata were considered: households with or without "migrants", without distinguishing current and return migrants. The sampling rate was higher in strata of households with migrants (return or abroad), in order to get a sufficient sample of such households. The selected number of households was 1920 in Ghana, 1773 in DR Congo, and 1320 in Senegal. The number of households successfully interviewed was 1246 in Ghana, 1577 in DR Congo, and 1141 in Senegal (Table 1).

In each of the selected households, one or several respondents were selected among the eligible people (people aged between 25 and 75, and born in the origin country[1]). The stratification at this stage was identical in the three countries. Three strata were distinguished: return migrants, partners of current migrants, and other household members. All the return migrants and partners of migrants currently abroad were selected. In addition, one other eligible member was randomly selected. In Senegal, the household survey and the individual survey were conducted subsequently. In Ghana and DR Congo, a special tool had been designed so that the interviewers could randomly select the people during the fieldwork (see Annex 1). Thus, the number of individuals could not be determined precisely before the survey, because it depended on the number of migrants and partners found in the households. In the end, the number of individuals successfully interviewed is a little higher than the number of households (1282 in Ghana, 1062 in Senegal and 1711 in DR Congo).

Table 1. Sampling characteristics in Senegal, Ghana, and DR Congo

	Senegal	Ghana	DR Congo
Target areas	Dakar Region (26% of the population of the country)	Accra and Kumasi (12% of the population of the country)	Kinshasa (17% of the population of the country)
Sampling frames of PSUs	2002 Population and Housing Census	2000 Population and Housing Census	Sampling frame of the 2007 DHS
1st stage: selection of primary sampling units	Selection of 60 census enumeration areas out of 2109, 6 in each stratum	Selection of 60 census enumeration areas in Accra and 20 in Kumasi	Selection of 29 neighbourhoods and 3 streets per neighbourhood (87 sampling units)
Stratification at 1st stage	Census districts were divided into 10 strata of equal size (equal number of districts) based on the migration prevalence (number of households with at least one migrant) in the district.	Two cities (Accra and Kumasi).	3 strata based on prevalence of migration

[1] In Senegal, an additional condition was that people had the Senegalese citizenship at birth. This condition was dropped in Ghana and DR Congo, because it complicated the sampling of individuals, and very few people born and living in these countries did not have the citizenship of the country at birth.

Table 1 (*cont'd.*). Sampling characteristics in Senegal, Ghana, and DR Congo

	Senegal	Ghana	DR Congo
2nd stage: selection of households	Enumeration to update household list. Random selection of 22 households per enumeration area. 11households selected in each of the two strata. If less than 11 households available in one or several strata, the remaining households are selected in the other stratum.	Enumeration to update household list. Random selection of 24 households per enumeration area. 8 households selected in each of the 3 strata. If less than 8 households available in one or several strata, the remaining households are selected in the other stratum.	Enumeration to update household list. Random selection of 21 households per enumeration area. 87households selected in each of the 3 strata. If less than 7 households available in one or several strata, the remaining households are selected in the other stratum. In a few streets, there were less than 21 households; all of them were selected.
Stratification at 2nd stage	2 strata households with and without migrants	3 strata: households with migrants abroad, with return migrants, without migrants	3 strata: households with migrants abroad, with return migrants, without migrants
3rd stage: selection of individuals	People aged 25-75, born in Senegal and who have/had Senegalese citizenship. Up to two return migrants and partners of migrants, and one randomly selected other eligible person.	People aged 25-75, born in Ghana. All the return migrants and partners of migrants, and one randomly selected other eligible person.	People aged 25-75, born in DR Congo. All the return migrants and partners of migrants, and one randomly selected other eligible person.
Stratification at third stage	3 strata: returnees, partners left behind and other non-migrants within households sampled at second stage	3 strata: returnees, partners left behind and other non-migrants within households sampled at second stage	3 strata: returnees, partners left behind and other non-migrants within households sampled at second stage
Sample size (selected households)	1320 households	1920 households (1440 in Accra and 480 in Kumasi)	1773 households
Completed household questionnaires*	1141 households, including: Non-migrant HH: 458 HH with at least 1 returnee: 205 HH with at least 1 current migrant: 617 Household with returnee(s) and current migrant(s): 139	1246 households, including Non-migrant HH: 449 HH at least 1 returnee: 346 HH with at least 1 current migrant:675 Household with returnee(s) and current migrant(s): 224	1576 households, including Non-migrant HH: 470 HH at least 1 returnee: 351 HH at least 1 current migrant:1027 Household with returnee(s) and current migrant(s): 272
Sample size (selected individuals)	1387	1490	1946
Completed biographic questionnaires	1062 individuals, including: Returnees: 193 Partners left behind: 101 Other non-migrants: 768	1243 individuals, including: Returnees: 319 Partners left behind: 84 Other non-migrants: 840	1638 individuals, including: Returnees:322 Partners left behind: 77 Other non-migrants: 1239
Household response rate	86.4 %	64.9 %	88.9 %
Individual response rate	76.6 %	83.4 %	84.2 %
Overall response rate	66.1%	54.1 %	74.9 %

COMPUTATION OF WEIGHTS IN AFRICA

The computation of sampling weights relies on computing sampling probabilities at each stage. The product of sampling probabilities at each stage gives the overall sampling probability. Taking the

inverse of the sampling probability gives the inflation factor. These factors are adjusted (trimming, adjusting for population size). They are then normalized, so that the sum of weights is equal to the sample size. The normalization of sampling weights depends on the type of analysis. It is explained in another note[2].

SAMPLE DESIGN OF THE GHANA MAFE SURVEY

The population covered in the MAFE survey is defined as the universe of all private households in Accra and Kumasi (the two largest cities of Ghana), as well as all adults aged 25-75 at the time of the survey. The sample of households is a stratified two-stage random sample.

SAMPLING FRAME

The Ghana MAFE survey used the list of census enumeration areas (EAs) with population and household information of the 2000 Population census as a sampling frame at the first stage. A listing operation was carried out in each of the 80 selected EAs to prepare the sampling frame of households. The listing operation consisted in enumerating all the households in the selected sites, and in identifying the 'migration status' of the household. Three categories of households were distinguished during the listing (households with return migrants, with migrants abroad, and without migrants), and constituted strata for the selection of households. Within each selected household, a list of eligible respondents was prepared, and the selection of individuals was done during the fieldwork.

SAMPLE SELECTION OF HOUSEHOLDS

Two strata were distinguished at the first stage, corresponding to the two cities covered by the survey. A total of 80 enumeration areas were selected, with a probability proportional to their estimated size. 60 EAs were selected in Accra, and 20 EAs in Kumasi.

At the second stage, 24 households were randomly selected in each of the 80 EAs. In total, 1920 households were selected. A stratification was done at the second stage, and a higher sampling rate was set for households with migrants (return or abroad), in order to get a sufficient sample of such households. In practice, the 24 households were selected in the following way: 8 households with return migrants were selected, then 8 households with migrants abroad, and finally 8 households without migrants. If less than 8 households were available in one or several strata, the remaining households were selected in the next strata. For instance, if only 4 household with return migrants were found in an EA, the 20 remaining other households were selected among household with migrants and without migrants. If only 6 households with migrants were found, all of them were included, and 14 households without migrants were selected. The stata syntax used for selecting households in Accra is presented in Annex 2.

SAMPLE SELECTION OF INDIVIDUAL RESPONDENTS

In each of the selected households, one or several respondents were selected among the eligible people (people aged between 25 and 75, and born in the origin country). All the return migrants and partners of migrants currently abroad were selected. In addition, one other eligible member was randomly selected. A special tool had been designed (based on DHS surveys) so that the interviewers could randomly select the people during the fieldwork (see example of selection grid in annex 1). The number of sampled individuals could not be determined precisely before the survey,

[2] Schoumaker B. (2011), *Note on the computation of normalized weights for the MAFE surveys*, Technical report, MAFE Project.

because it depended on the number of migrants and partners found in the households. In the end, the number of individuals successfully interviewed is a close to the number of households.

RESPONSE RATES FOR HOUSEHOLD AND BIOGRAPHIC SURVEYS

Information on household and individual interviews is presented in Table 2. A total of 1920 households were selected for the MAFE survey. Household interviews were completed for 1246 households, giving a response rate of 65% on average for the household survey. A total of 1490 respondents were selected for the biographic survey (out of 2315 eligible respondents)[3]. Biographic surveys were completed for 1243 individuals (response rate of 83% on average). The overall response rate is obtained by multiplying the household response rate and the individual response rate. It is around 54% on average, but varies strongly between Accra and Kumasi. In Accra, the overall response rate is a little below 50%, whereas it is over 75% in Kumasi. The non-response rate is higher in Ghana than in the other MAFE countries.

Table 2. Number of primary and secondary sampling units, and number of households and individuals by results of the household and individual interviews, according to strata (Ghana).

	Strata		TOTAL
	Accra	Kumasi	
Number of PSU (EAs)	60	20	29
Number of selected households	1440	480	1920
Number of completed household surveys	866	380	1246
Household response rate	60.1%	79.2%	64.9%
Number of eligible respondents	1618	697	2315
Number of selected respondents	1017	473	1490
Number of completed individual questionnaires	794	449	1243
Selected respondents response rate	78.1%	94.9%	83.4%
Overall response rate	47.0%	75.1%	54.1%

COMPUTATION OF WEIGHTS IN THE GHANA MAFE SURVEY

The computation of sampling weights relies on computing sampling probabilities at each stage. The product of sampling probabilities at each stage gives the overall sampling probability. Taking the inverse of the sampling probability gives the inflation factor. These factors are adjusted (trimming, adjusting for population size). They are normalized, so that their sum is equal to the sample size.

SAMPLING PROBABILITIES FOR HOUSEHOLDS

The first stage of sampling was done by selecting EA systematically with probability proportional to estimated size. In a stratum (Accra or Kumasi), the selection probability of EA j is computed as

$$P_{1j} = a * \frac{H_j}{\sum H_j}$$

[3] The number of selected respondents is lower than the number of eligible persons. All return migrants and partners of migrants were selected, but only one other eligible person could be selected per household.

Where a is the number of selected EAs in the strata, H_j is the number of households in the j[th] EA according to the 2000 Population Census, and $\sum H_j$ is the number of households in the stratum according to the Census.

In each selected EA, a listing of households was carried out, and listed households were classified into three substrata (non migrants, with return migrants, with migrants abroad). At the second stage, households were selected within each EA with varying probabilities across the three strata.

$$P_{2jk} = \frac{S_{jk}}{M_{jk}}$$

Where M_{jk} is the number of households in substrata k listed in the j[th] EA, and S_{jk} is the number of selected household in substrata k in the j[th] EA. As explained before, households with return migrants and migrants abroad were oversampled, so P_2s are higher for these types of households.

Non response was also taken into account in the computation of weights. To do this, sampling probabilities at the second stage were actually computed by dividing the number of *completed interviews* (instead of selected households) by the number of households.

$$P^*_{2jk} = \frac{C_{jk}}{M_{jk}}$$

Where C_{jk} is the number of completed household interviews in substrata k in the j[th] EA.

The overall sampling probability for households is the product of the sampling probabilities at the first and second stages.

$$f_h = P_{1j}.P^*_{2jk}$$

The inflation factor for households is computed as the inverse of the overall sampling probability.

$$w_h = 1/f_h$$

TRIMMING OF HOUSEHOLD WEIGHTS

There is a trade-off to the decision whether to trim, i.e. remove, extremely large or small weights or not. On the one hand, weights should not be modified in order to maintain their role in eliminating bias due to unequal selection probabilities of cases. On the other hand, the introduction of extreme weights, even if they affect only a small number of cases, can increase considerably the variance of estimates. Because weights vary largely across households, it was decided to trim the weights in order to limit to 100 the ratio between the maximum weight and the minimum weight (see the Stata syntax on Ghana, annex 3). Trimming reduces the impact of weighting on sampling variance, but may lead to small biases. The trimmed weights are noted w^*_h.

ADJUSTING TO TOTAL NUMBER OF HOUSEHOLDS

The sum of weights is supposed to be equal to the total number of households. For various reasons (problems in the listing phase, trimming of weights…), the sum of weights may differ in practice from the total number of households. An adjustment factor was computed as the ratio of the total number of households estimated in 2009 (United Nations, 2009) in the population and the sum of weights. The adjusted trimmed household weights are noted w^{**}_h.

INDIVIDUAL WEIGHTS: SAMPLING PROBABILITIES, TRIMMING, ADJUSTMENT

At the third stage, individuals were selected in each household among the eligible respondents that were divided into three substrata (non migrants, return migrants, partner of migrants). Only a subset of the selected respondents participated in the survey. Sampling and response rates are combined in the following way, in order to compute probabilities of selection at the third stage.

$$P_{3l}^{*} = \frac{R_l}{N_l}.$$

Where R_l is the number of completed individual interviews among members in substratum l in the household and N_l is the number of eligible individuals in substrata l in the household[4].

The overall sampling probability for individuals is the product of the inverse of the adjusted household weights, and of the sampling probability at the individual level.

$$g_b = \frac{P_{3l}^{*}}{w_h^{**}}$$

The inflation factor for the individual (biography) is computed as the inverse of the overall sampling probability.

$$w_b = 1/g_b$$

Individual weights are also trimmed in order to limit to 100 the ratio between the maximum weight and the minimum weight. The trimmed individual weights are noted w_b^{**}.

SAMPLE DESIGN OF THE DR CONGO MAFE SURVEY

The population covered in the MAFE survey is defined as the universe of all private households in Kinshasa (the capital city of DR Congo), as well as all adults aged 25-75 at the time of the survey. The sample of households is a stratified three-stage random sample.

SAMPLING FRAME

The Congo MAFE survey used the sampling frame of primary sampling units (324 neighbourhoods) prepared for the 2007 DHS. The neighbourhoods were not small enough for a complete household listing. It was therefore necessary to subdivide each neighbourhood into smaller units. For each selected neighbourhood, a list of the streets and the estimated number of plots was obtained from the "chefs de quartiers". Three streets (secondary sampling units) were the selected in each of the selected primary sampling units. The listing consisted in enumerating all the households in the selected sites, and in identifying the 'migration status' of the household. Three categories of households were distinguished during the listing (households with return migrants, with migrants abroad, and without migrants), and constituted strata for the selection of households. Within each selected household, a list of eligible respondents was prepared, and the selection was done during the fieldwork.

SAMPLE SELECTION OF HOUSEHOLDS

The 324 neighbourhoods (primary sampling units) were divided into three strata, based on the prevalence of migration. The prevalence of migration was estimated from expert knowledge: 6

[4] For the sake of clarity, subscripts are omitted, but sampling rates vary across households.

Congolese migration specialists classified each of the 24 communes into three strata (high, medium and low migration).

At the first stage, 29 of the 324 neighbourhoods (primary sampling units, PSUs) were selected with a probability proportional to their estimated size. The sampling rate was highest in the high migration stratum, and lowest in the low migration stratum[5].

At the second stage, 3 secondary sampling units (streets) were also selected with a probability proportional to their estimated size (number of plots) in each of the 29 neighbourhoods. In total, 87 streets were selected in the sample.

At the third stage, 21 households were randomly selected in each of the 87 streets, except in a few small streets where the number of households was smaller than 21. In total, 1773 households were selected (on average, 20.4 households per street). A stratification was done at the third stage, and a higher sampling rate was set for households with migrants (return or abroad), in order to get a sufficient sample of such households. In practice, the 21 households were selected in the following way: 7 households with return migrants were selected, then 7 households with migrants abroad, and finally 7 households without migrants. If less than 7 households were available in one or several strata, the remaining households were selected in the other strata. For instance, if only 4 household with return migrants were found in an EA, the 17 remaining other households were selected among household with migrants and without migrants. If only 5 households with migrants were found, all of them were included, and 12 households without migrants were selected (the stata syntax used for selecting households in Accra is presented in Annex 2; a similar syntax was used in Kinshasa).

SAMPLE SELECTION OF INDIVIDUAL RESPONDENTS

In each of the selected households, one or several respondents were selected among the eligible people (people aged between 25 and 75, and born in the origin country). All the return migrants and partners of migrants currently abroad were selected. In addition, one other eligible member was randomly selected. A special tool had been designed so that the interviewers could randomly select the people during the fieldwork (see example in annex for Ghana). The number of sampled individuals could not be determined precisely before the survey, because it depended on the number of migrants and partners found in the households. In the end, the number of individuals successfully interviewed is a little higher than the number of households.

RESPONSE RATES FOR HOUSEHOLD AND BIOGRAPHIC SURVEYS

Information on household and individual interviews is presented in Table 3. A total of 1773 households were selected for the MAFE survey. Household interviews were completed for 1577 households, giving a response rate of 89% on average for the household survey. A total of 1946 respondents were selected for the biographic survey (out of 4238 eligible respondents)[6]. Biographic surveys were completed for 1638 individuals (response rate of 84% on average). The overall response rate is obtained by multiplying the household response rate and the individual response rate. It is around 75% on average, and does not vary across strata. These response rates are overall quite satisfying, and higher than in Ghana and Senegal.

[5] 17 were selected in the high migration strata, 6 in the medium migration strata, and 6 in the low migration stratum.

[6] The number of selected respondents is lower than the number of eligible persons. All return migrants and partners of migrants were selected, but only one other eligible person was selected per household.

Table 3. Number of primary and secondary sampling units, and number of households and individuals by results of the household and individual interviews, according to strata (DR Congo).

	Strata			TOTAL
	High	Medium	Low	
Number of PSU (neighbourhoods)	17	6	6	29
Number of SSU (streets)	51	18	18	87
Number of selected households	1053	346	374	1773
Number of completed household surveys	949	296	331	1577
Household response rate	90.1%	85.5 %	88.5 %	88.9 %
Number of eligible respondents	2558	787	893	4238
Number of selected respondents	1177	346	423	1946
Number of completed individual questionnaires	974	301	361	1638
Selected respondents response rate	82.8%	87.0%	85.3%	84.2%
Overall response rate	74.6%	74.4%	75.5%	74.9%

COMPUTATION OF WEIGHTS IN THE DR CONGO MAFE SURVEY

The computation of sampling weights relies on computing sampling probabilities at each stage. The product of sampling probabilities at each stage gives the overall sampling probability. Taking the inverse of the sampling probability gives the inflation factor. These factors are adjusted (trimming, adjusting for population size). They are normalized, so that their sum is equal to the sample size.

SAMPLING PROBABILITIES FOR HOUSEHOLDS

The first stage of sampling was done by selecting neighbourhoods systematically with probability proportional to estimated size. In each stratum (one of the three strata at first stage), the selection probability of neighbourhood i is computed as

$$P_{1i} = a * \frac{H_i}{\sum H_i}$$

Where a is the number of selected EAs in the strata, H_i is the number of households in the ith EA according to the sampling frame, and $\sum H_i$ is the number of households in the stratum according to the 2007 DHS sampling frame.

In each selected EA, a list of streets with the number of plots was used to randomly select three streets, with a probability proportional to the number of plots.

$$P_{2ij} = 3 * \frac{L_{ij}}{\sum L_{ij}}$$

Where L_{ij} is the number of plots in the jth street of neighbourhood i, and $\sum L_{ij}$ is the number of plots in the neighbourhood i.

In each selected street, a listing of household was carried out, and listed households were classified into three substrata (non migrants, with return migrants, with migrants abroad). At the second

stage, households were selected within each street with varying probabilities across the three strata.

$$P_{3ijk} = \frac{S_{ijk}}{M_{ijk}}$$

Where M_{hjk} is the number of household in substrata k listed in the j[th] street of neighbourhood i, and S_{ijk} is the number of selected households in substrata k in the j[th] street of neighbourhood i. As explained before, household with return migrants and migrants abroad were oversampled, so P_3s are higher for these types of households.

Non-response was also taken into account in the computation of weights. To do this, sampling probabilities at the second stage were actually computed by dividing the number of *completed interviews* (instead of selected households) by the number of households.

$$P^*_{3ijk} = \frac{C_{ijk}}{M_{ijk}}$$

Where C_{ijk} is the number of completed household interviews in substrata k in the j[th] street of neighbourhood i.

The overall sampling probability for households is the product of the sampling probabilities at the first, second and third stages.

$$f_h = P_{1i} . P_{2ij} . P^*_{3ijk}$$

The inflation factor for households is computed as the inverse of the overall sampling probability.

$$w_h = 1/f_h$$

TRIMMING OF HOUSEHOLD WEIGHT

Because weights vary largely across individual observations, it was decided to trim the weights in order to limit the ratio between the maximum weight and the minimum weight to 100. Trimming reduces the impact of weighting on sampling variance, but may lead to small biases. The trimmed weights are noted w^*_h.

ADJUSTING TO TOTAL NUMBER OF HOUSEHOLDS

The sum of weights is supposed to be equal to the total number of households. For various reasons (problems in the listing phase, trimming of weights…), the sum of weights may differ in practice from the total number of households. An adjustment factor was computed as the ratio of the total number of households estimated in 2009 (United Nations, 2009) in the population and the sum of weights. The adjusted trimmed household weights are noted w^{**}_h.

INDIVIDUAL WEIGHTS: SAMPLING PROBABILITIES, TRIMMING, ADJUSTMENT

At the third stage, individuals were selected in each household among the eligible respondents that were divided into three substrata (non migrants, return migrants, partner of migrants). Only a subset of the selected respondents participated in the survey. Sampling and response rates are combined in the following way, in order to compute probabilities of selection at the fourth stage.

$$P^*_{4l} = \frac{R_l}{N_l}.$$

Where R_l is the number of completed individual interviews among members in substratum l in the household and N_l is the number of eligible individuals in substrata l in the household[7].

The overall sampling probability for individuals is the product of the inverse of the adjusted household weights, and of the sampling probability at the individual level.

$$g_b = \frac{P_{il}^*}{w_h^{**}}$$

The inflation factor for the individual (biography) is computed as the inverse of the overall sampling probability.

$$w_b = 1/g_b$$

Individual weights are also trimmed in order to limit the ratio between the maximum weight and the minimum weight to 100. The adjusted trimmed individual weights are noted w_b^{**}.

SAMPLE DESIGN OF THE SENEGAL MAFE SURVEY

The population covered in the MAFE survey is defined as the universe of all private households in the region of Dakar (household survey), as well as all adults aged 25-75 at the time of the survey who were born in Senegal and had Senegalese citizenship at birth (individual survey). The sample of households is a stratified two-stage random sample.

SAMPLING FRAME

The Senegal MAFE survey used the list of census enumeration areas (EAs, also called 'census districts') with population and household information of the 2002 Population census as a sampling frame at the first stage. A listing operation was carried out in each of the 60 selected EAs to prepare the sampling frame of households. The listing operation consisted in enumerating all the households in the selected sites, and in identifying the "migration status" of the household. Two categories of households were distinguished during the listing (households with migrants and without migrants), and constituted strata for the selection of households.[8] After completion of the household survey, a list of eligible individual respondents was prepared, from which a sample of respondents of the individual survey was drawn.

SAMPLE SELECTION OF HOUSEHOLDS

At the first stage, the 2 109 districts in the Dakar region were ranked according to the proportion of households that had declared to have one or several migrants abroad in the 2002 census. Districts were then divided into 10 strata of equal size (9 strata with 211 districts and 1 stratum with 210 districts). A total of 60 districts were selected (6 per stratum). The selection probability of a census district within a given stratum was proportional to its size in terms of the number of households residing in the district at the time of the census in 2002.

At the second stage, 22 households were randomly selected in each of the 60 districts sampled at the first stage. Households were divided in two strata, households with and households without migrants, in order to obtain a sufficient sample of households with migration experience. In

[7] For the sake of clarity, subscripts are omitted, but sampling rates vary across households.

[8] During the enumeration phase, the following question were asked to all the households: "does your household include one or more migrants?", with only two possible answer categories (yes or no).

general, 11 households were selected randomly in each stratum. However, if less than 11 households were available in one stratum, the remaining households were selected in the other stratum in order to achieve a total number of 22 households per district. For instance, if only 4 households with migrants were found in a district, all of them were selected, and the 18 remaining households were selected among the households without migrants. A total of 1 320 households (449 with migrants and 841 without) constituted the household sample.

SAMPLE SELECTION OF INDIVIDUAL RESPONDENTS

In each of the selected households, one or several respondents were selected among the eligible individuals (people aged between 25 and 75, and born in the origin country[9]). All the return migrants and partners of migrants currently abroad were selected. In addition, one other eligible member was randomly selected. The random selection was made by computer, using a file obtained after data entry of key variables from the household questionnaire that were required to determine eligibility for the biographic survey.

RESPONSE RATES FOR HOUSEHOLD AND BIOGRAPHIC SURVEYS

Information on response rates for both household and individual interviews is presented in Table 4. A total of 1320 households were selected for the MAFE survey. Household interviews were completed for 1141 households, giving a response rate of 86% on average for the household survey (86% for households with migrants and 89% for those without migrants). A total of 1393 individuals were then selected for the biographic survey (out of 4185 eligible persons).[10] Biographic surveys were completed for 1062 individuals (response rate of 76% on average). Depending on the stratum, the household response rate varied between 75.7% and 92.4%. The overall response rate is obtained by multiplying the household response rate and the individual response rate. It is around 65% on average, but varies across the strata, from 58% in stratum 10 (the stratum with the highest proportion of migrant households according to the census) to 74% in stratum 4.

Table 4. Number of primary and secondary sampling units, and number of households and individuals by results of the household and individual interviews, according to strata (Senegal).

	Strata										Total
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	
Number of districts (PSUs)	6	6	6	6	6	6	6	6	6	6	60
Number of selected households (SSUs)	132	132	132	132	132	132	132	132	132	132	1320
Number of completed household surveys	116	122	112	115	112	117	115	100	116	116	1141
Household response rate	**87.9%**	**92.4%**	**84.9%**	**87.1%**	**84.9%**	**88.6%**	**87.1%**	**75.8%**	**87.9%**	**87.9%**	**86.4%**
Number of eligible individual respondents	423	438	424	330	362	453	504	370	445	436	4185
Number of selected respondents	138	139	143	127	132	136	151	127	154	146	1393
Number of completed individual questionnaires	100	102	103	108	103	109	115	100	126	96	1062
Selected respondents response rate	**72.5%**	**73.4%**	**72.0%**	**85.0%**	**78.0%**	**80.2%**	**76.2%**	**78.7%**	**81.8%**	**65.8%**	**76.2%**
Overall response rate	**63.7%**	**67.8%**	**61.1%**	**74.2%**	**66.2%**	**71.0%**	**66.4%**	**59.7%**	**71.9%**	**57.8%**	**65.9%**

[9] In Senegal, an additional condition was that people had the Senegalese citizenship at birth. This condition was dropped in the Ghana and DR Congo surveys, because it complicated the sampling of individuals, and very few people born and living in these countries did not have the citizenship of the country at birth.

[10] The number of selected individuals is lower than the number of eligible persons. All return migrants and partners of migrants were selected, but only one other eligible person could be selected per household.

COMPUTATION OF WEIGHTS IN THE SENEGAL MAFE SURVEY

The computation of sampling weights (design weights) relies on computing sampling probabilities at each stage. The product of sampling probabilities at each stage gives the overall sampling probability. Taking the inverse of the sampling probability gives the inflation factor. These factors are adjusted (taking into account non-response and by trimming the weights). In a final step, the weights are normalized, so that their sum is equal to the sample size.

SAMPLING PROBABILITIES FOR HOUSEHOLDS AND INITIAL INFLATION WEIGHTS

The first stage of sampling was done by selecting districts systematically with probability proportional to estimated size. In each stratum, the selection probability of EA j is computed as

$$P_{1j} = a * \frac{H_j}{\sum H_j}$$

Where a is the number of selected EAs in the stratum, H_j is the number of households in the j[th] EA according to the 2002 Census, and $\sum H_j$ is the number of households in the stratum according to the Census.

In each selected EA, a listing of households was carried out, and listed households were classified into two substrata (households without migrants, with migrants). At the second stage, households were selected with varying probabilities across the two strata within each EA:

$$P_{2jk} = \frac{S_{jk}}{M_{jk}}$$

where M_{jk} is the number of households in substratum k listed in the j[th] EA, and S_{jk} is the number of selected eligible households in substrata k in the j[th] EA. As explained before, households with migrants were oversampled, so sampling probabilities P_2 are, at the average, higher for these types of households.

The overall household sampling probabilities are computed as the product of sampling probabilities at the first and second stages:

$$f_h = P_{1j} . P_{2jk}$$

The initial design weight (inflation factor) w_h^{in} is computed as the inverse of the overall sampling probability:

$$w_h^{in} = 1/f_h$$

NON-RESPONSE ADJUSTMENT OF HOUSEHOLD WEIGHTS

Unit non-response, meaning that a fraction of the sampled households and/or individuals does not respond at all to the questionnaire, is a possible source of nonsampling error.[11] If non-response is

[11] Other sources of nonsampling error include noncoverage (incomplete sampling frames) as well as observation errors when observations are incorrectly obtained in the field (e.g. interviewer bias) or processed (e.g. data entry). Noncoverage at the level of primary sampling units is less likely when the frame is based on census or large-scale national survey data and primary sampling units are census districts or neighborhoods. Observation errors are not accounted for in the current versions of the MAFE data. However, interview and interviewer characteristics have been collected and recorded in the databases and could be used by interested data users to explore observational errors to some extent.

not random but differs depending on characteristics of the household or individual, biases may be introduced when analyzing the realized sample. Weighted response rates are computed as the sum of initial design weights w_h^{in} of households that responded divided by the sum of initial design weights w_h^{in} of eligible households in census stratum j and household stratum k:

$$R_{jk}^{h} = \frac{\sum(w_h^{in})_{jk}^{R}}{\sum(w_h^{in})_{jk}^{E}}$$

The non-response adjusted household inflation weight is equal to:

$$w_h = w_h^{in} / R_{jk}^{h}$$

TRIMMING OF HOUSEHOLD WEIGHTS

Household sampling weights were not trimmed, because the ratio of extreme weights was lower than 100.

SAMPLING PROBABILITIES FOR INDIVIDUALS AND INITIAL INFLATION WEIGHTS

At the third stage, individuals were selected in each household among the eligible respondents that were divided into three substrata (non migrants, return migrants, partner of migrants[12]).

The selection probability for an individual within stratum l is thus equal to:

$$P_{il} = \frac{I_l}{N_l}$$

where I_l is the number of selected individuals in stratum l in a given household and N_l is the number of individuals eligible in stratum l in a given household. The overall initial design weight at the individual level is the product of the weight at household level and at individual level:

$$w_b^{in} = w_h . 1/P_{il}$$

NON-RESPONSE ADJUSTMENT OF INDIVIDUAL WEIGHTS

As shown in Table 4, only a subset of the selected respondents participated in the survey. In the Senegalese case, the main reason for non-response at the individual level was due to noncontact, despite the fact that interviewers tried to contact selected individuals up to ten times (Razafindratsima et al., 2010). Non-response adjustments were computed based on response propensities within homogeneous response groups. Thanks to data collected in the household survey, more information has been available to compute response propensities at the individual level. In addition to the stratum (non-migrant, spouse of migrant, or return migrant), age, sex, household size and the number of contacts have been used as explanatory variables in a logistic regression. Individuals were ranked by their estimated response propensity and divided into homogenous groups *g*.

[12] Partners of migrants who were return migrants were assigned to the return migrant stratum.

The response rate was computed as the sum of initial design weights w_b^{in} of individuals who responded divided by the sum of initial design weights w_b^{in} of eligible individuals within each homogeneous group *g*:

$$R_g^b = \frac{\Sigma\left(w_b^{in}\right)_g^R}{\Sigma\left(w_b^{in}\right)_g^E}$$

The non-response adjusted individual inflation weight is equal to:

$$w_b = w_b^{in} / R_g^b$$

TRIMMING OF INDIVIDUAL WEIGHTS

There is a trade-off to the decision whether to trim, i.e. remove, extremely large or small weights or not. On the one hand, weights should not be modified in order to maintain their role in eliminating bias due to unequal selection probabilities of cases. On the other hand, the introduction of extreme weights, even if they affect only a small number of cases, can increase considerably the variance of estimates. In the case of the Senegalese individual data, the initial ratio between maximum and minimum weights was approximately 220, indicating a very high level of dispersion. It was therefore decided to trim the extreme weights to reduce the ratio to approximately 100 while limiting the number of cases affected. The 18 cases with the highest and the 18 cases with the lowest weight were truncated (3.4% of all individuals). The steps involved (i) assigning the cut-off weights after truncation to individuals whose weights were below or above the cut-offs and (ii) scaling the remaining weights so that the sum of all weights corresponded again to the estimate of the population size. In this way, the ratio between maximum and minimum weights was reduced to 97.4.

The trimmed individual weights are noted w_b^{**}.

VARIANCE ADJUSTMENT

Sampling weights correct the bias in point estimates due to differential representation of sample observations. However, sampling designs (with stratification with disproportionate sampling fractions and clustering) may also affect the variance estimation. Stratification usually leads to smaller variances than in simple random sampling if units within strata are homogenous, while clustering entails a larger variance, since observations in the same cluster tend to be correlated. The standard error estimates may vary from “only weights” to “weights + design features”, especially if clustering is not accounted for. Standard errors may also be affected in case of disproportionate stratification when certain population groups are oversampled.

Most statistics software programmes include now commands that allow for correction of the variance estimation, in addition to the inclusion of weights. In order to do so, the MAFE databases for Senegal, Ghana, and Congo also include variables to identify the clusters and strata. Household data includes information about levels 1 and 2, while individual data includes information about all three levels. In most cases, adjustments only take account of sample selection at level 1.

Table5. Design variables included in MAFE datasets for African samples

STAGE	VARIABLE NAME	VARIABLE DESCRIPTION
Level 1 Survey Area	num_dr; nodr...	In Senegal and Ghana, census districts are primary sampling units In Congo, neighbourhoods are primary sampling units
	strata_area	- In Senegal, census districts are divided into 10 strata according to migration prevalence information from the 2002 Census - In Congo, neighbourhoods are divided into three strata based on migration prevalence information, defined through interviews with knowledgeable persons (1= High level of migration ; 2 = Medium level ; 3= low level) - In Ghana, the two survey regions constitute strata at the highest level (1=Accra ; 2=Kumasi)
Level 2 Household	n_menage	Secondary sampling units are households
	strata_hh	Presence of migrant in the HH during the enumeration. - in Senegal : 0=non migrant HH ; 1= HH with returnee(s) or migrant(s) abroad - in Ghana and Congo : 0=HH with no migrant ; 1=HH with returnee(s) ; 2=HH with current migrants *Note: the numbers reported may be different from what was actually observed in the HH data as definition of strata was based on a simple question during the enumeration (see the following variables: HMIGTOT and HMIGRET)*
Level 3 Individual	Ident	Tertiary sampling units are individuals
	strata_ind	Strata of the individuals draw in 3 classes. In all African countries, the categories are the following: 1 = return migrants ; 2= spouse of a migrant ; 3 = Other non migrant *Note: the numbers reported may be different from what was actually observed in the individual data (see the following variables: migr_cur, migr_ret, migr_cjt, migr_no)*

SAMPLING AND COMPUTATION OF WEIGHTS IN EUROPE

BRIEF DESCRIPTION OF SAMPLING IN EUROPE

With the exception of Spain, no sampling frames existed in European countries (France, Italy, Netherlands, United Kingdom, and Belgium) to draw probabilistic samples of Senegalese, Congolese and Ghanaian migrants (Beauchemin, 2012; Schoumaker and Diagne, 2010). Instead, quota sampling methods were used to select the European samples of the MAFE surveys. Quotas were established by age and sex, and additional characteristics (socio-economic status, place of residence) were used in some of the European destination countries (Table 3). Various recruiting methods were used (in public spaces, volunteer lists at churches, associations, contacts obtained from origin households in the case of MAFE-Senegal etc.) and the sampling composition was monitored during fieldwork to make sure that quotas were filled, but also to ensure a diversity with regard to other characteristics, such as education or region of origin.

Table 6. Sampling approaches in European countries

Country	Target areas	Sample size	Quotas	Recruitment methods
MAFE-Senegal				
France	3 selected regions: Ile de France, around Paris; Rhône-Alpes, around Lyon; Provence-Alpes-Côte d'Azur, around Marseille.	201 (46% of females), including undocumented migrants - at the time of the survey: 12%(1) - in the past(2): 29% 80 % have lived at least one year in the region of Dakar	By age, gender and socio-economic status	Selection from contacts obtained in Senegal, Public spaces, migrant associations, snowballing, interviewers' contacts
Italy	4 selected regions: Lombardia, Emilia Romagna, Toscana, Campania.	205 (39% of females), including undocumented migrants - at the time of the survey: 17% - in the past: 46% 54% have lived at least one year in the region of Dakar	By age and gender	Selection from contacts obtained in Senegal, Public spaces, migrant associations, snowballing, interviewers' contacts
Spain	12 provinces: Almería (Andalucía); Alicante & Valencia (Comunidad Valenciana); Barcelona, Lérida, Tarragona & Gerona (Cataluña) ; Madrid (Comunidad de Madrid); Zaragoza (Aragón); Las Palmas (Islas Canarias); Murcia (Comunidad Autónoma de Murcia) ; Baleares (Islas Baleares)	200 (51% of females), including undocumented migrants - at the time of the survey: 18% - in the past: 57% 61 % have lived at least one year in the region of Dakar. NB: an additional sample of around 400 people will be added, thanks to a new survey round carried out in 2010.	Random sample from Padron	Population register (Padron) & contacts obtained in Senegal
MAFE-Congo				
Belgium	Whole country	279 (45% of females), including undocumented migrants - at the time of the survey: 10% - in the past: 33% 87.5 % have lived at least one year in Kinshasa	By age, gender and place of residence	Public spaces, migrant associations, churches, snowballing, phonebook, centers for asylum seekers, interviewers' contacts
United Kingdom	Whole country	149 (50% of females), including undocumented migrants - at the time of the survey: 12% - in the past: 52% 93.3 % have lived at least one year in Kinshasa	By age, gender and place of residence	Public spaces, churches, snowballing, interviewers' contacts
MAFE-Ghana				
The Netherlands	3 cities (in 3 different provinces): Amsterdam(North Holland); The Hague (South Holland); Almere (Flevoland)	273 (47% of females), including undocumented migrants - at the time of the survey: 19% - in the past: 56% 72.5% have lived at least one year in Accra or Kumasi areas	By age and gender	Public spaces, churches, snowballing, interviewers' contacts
United Kingdom	Whole country	149 (48% of females), including undocumented migrants - at the time of the survey: 7% - in the past:14% 79.2% have lived at least one year in Accra or Kumasi areas	By age, gender and place of residence	Public spaces, churches, snowballing, interviewers' contacts

Notes: (1) Non-weighted percentage of interviewees having declared that they did not hold a residence permit at the time of the survey. (2) Non-weighted percentage of interviewees having declared that they did not hold a residence permit at some point in their migrant life for a period of at least one year (i.e. at the time of the survey or sometime in the past when they were living out of their origin country).

BRIEF DESCRIPTION OF COMPUTATION OF WEIGHTS IN EUROPE

In the European countries, similar sample sizes were selected for males and females, resulting in an overrepresentation or underrepresentation in the MAFE samples. Similarly, older people were usually oversampled. For these reasons, post-stratification weights are computed to give each observation its proper weight and to match the samples as closely as possible to selected population characteristics.

The computation of weights in most European data sets[13] is done using the following steps:

1) First, a relative age-sex distribution of the migrant population in the destination country is estimated. Several sources, which vary from one country to another, may be used to estimate the distribution (census data, population register, labor force survey, MAFE data collected in the origin countries).
2) Post-stratification weights are computed by comparing the age-sex distribution of the migrants in the MAFE survey in the European country to the estimated age-sex distribution of the migrants in the country. For instance, the age-sex distribution of Ghanaians in the MAFE survey in the UK is compared to the estimated age-sex distribution of Ghanaian migrants in the UK. Post-stratification weights are equal to the ratio of the percentage in each cell in the population to the percentage in the MAFE survey.
3) Next, the size of the total migrant population from the origin country X (e.g. Ghanaians) in destination country Y (e.g. UK) is estimated. Again, several sources may be used to reach a reasonably accurate estimate of the size of the migrant population aged 25 and over.
4) Inflating factors are computed by multiplying the post-stratification weights by the ratio of the population size to the sample size.
5) Post-stratification weights (mean equal to 1) are used when a single survey is analyzed. Inflating factors are used when several surveys are pooled together.

COMPUTATION OF WEIGHTS FOR THE CONGOLESE SAMPLES IN EUROPE

BELGIUM – CONGOLESE MIGRANTS

Estimated age-sex distribution of Congolese migrants in Belgium

The age-sex distribution of Congolese migrants in Belgium is estimated from 2 data sources:

(a) The age-sex distribution of Congolese migrants from the population register in 2006[14].

(b) The age-sex distribution of Congolese migrants in Belgium identified in the MAFE survey conducted in Kinshasa (2009).

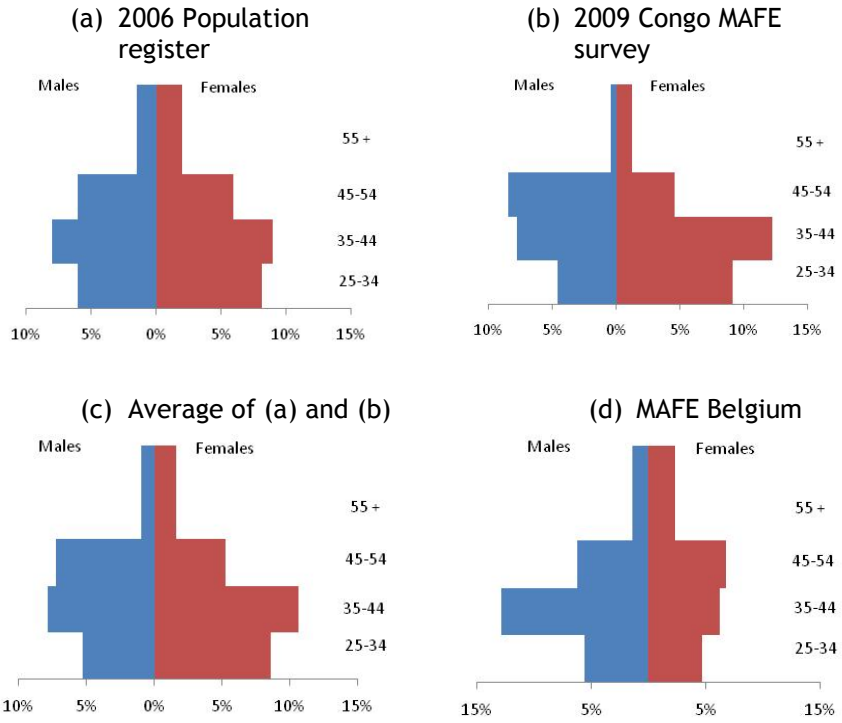
The relative age-sex distributions from these sources are represented on Figure 1(a) and 1(b). The estimated age-sex distribution of the Congolese migrant population in Belgium is estimated as the

[13] In the MAFE surveys in France, Spain, and Italy (Senegalese migrants), the order of steps was slightly different. Target population totals were estimated in a first step, followed by the computation of an initial inflation weight. Thirdly, post-stratification weights were computed (wither based on cell frequencies or on margins).

[14] Computations made by Quentin Schoonvaere, at Université catholique de Louvain.

average of the two distributions (Figure 1(c)). Although this is a rough approach, it gives a reasonable estimate of the age-sex structure. The distribution from the population register is based on more observations, but it does not represent perfectly the target population of the MAFE surveys: it includes migrants who arrived in Belgium before age 18 (not covered by MAFE surveys)[15], and undocumented migrants and asylum seekers are not included. The age-sex distribution of the MAFE surveys is affected by larger sampling errors, and only covers migrants living in Belgium who had relatives in Kinshasa; on the other hand, the data is more recent and also include undocumented migrants and asylum seekers, and it is possible to select people who migrated after age 18.

Figure 1(a) to 1(d). Age-sex structures of Congolese migrants in Belgium from various sources.



The gender distributions are relatively similar in the register and the MAFE Congo survey and indicate that women are a little more numerous than males among Congolese migrants in Belgium (57% of females in the MAFE-Congo survey, and 53% in the Population Register). The age structure, in contrast, is significantly different between sources, especially among females. However, according to both sources, around two-thirds of Congolese migrants are aged between 25 and 45.

Post-stratification weights

Figure 1(d) and the first part of Table 1 show the age-sex distribution of the respondents in the Congo MAFE survey in Belgium. Overall, males are overrepresented (55% compared to 45% in the average distribution), and older people are also (purposefully) overrepresented in the MAFE samples. Post stratification weights are computed as the ratios of percentages in the average distribution to the percentages in the MAFE Belgium survey. Cells with large post-stratification weights (e.g. females below 35) indicate that migrants with these characteristics were underrepresented in the MAFE survey.

[15] All tables refer to immigrated population born with Congolese citizenship (born abroad, i.e. not all individuals were necessarily born in DR Congo) who are older than 26 as of 1.1.2006.

Table 7. Relative age-sex distribution of Congolese migrants in the MAFE survey in Belgium, estimated age-sex distribution of Congolese migrants in Belgium, and post-stratification weights.

	MAFE Congo-Belgium			Estimated age-sex structure			Post-stratification weights	
Age groups	**Males**	**Females**	**Total**	**Males**	**Females**	**Total**	**Males**	**Females**
25-34	11.1%	9.3%	20.4%	10.6%	17.3%	27.9%	0.95	1.86
35-44	25.8%	12.5%	38.3%	15.8%	21.2%	37.0%	0.61	1.69
45-54	12.5%	13.6%	26.1%	14.4%	10.5%	24.9%	1.15	0.77
55 and over	5.7%	9.3%	15.0%	3.8%	6.4%	10.2%	0.66	0.69
Total	55.2%	44.8%	100%	44.6%	55.4%	100%		

Estimated population of Congolese migrants aged 25-75 in Belgium and inflating factors

The next step consists in estimating the population of Congolese migrants aged 25-75 in Belgium at the time of the survey. According to the Population Register, around 24 000 Congolese migrants aged 25-75 were living in Belgium in 2006. Extrapolating past trends of the population of Congolese migrants leads to about 27000 Congolese aged 25-75 in 2009 in Belgium. This is probably a low estimate – because undocumented migrants and asylum seekers are not included in the statistics. However, because no other reliable source exists, we use this as a sufficiently reasonable estimate of the Congolese migrant population in Belgium. The sampling rate in the MAFE Belgium survey is a little above 1/100.

Table 8 shows the age-sex structure in the MAFE sample, and the estimated numbers of Congolese migrants by age and sex in Belgium. The inflating factors are obtained by computing the ratio of estimated population by age and sex to the sample size by age and sex in the MAFE Belgium sample. The inflating factors are also equal to the post-stratification weights divided by the sampling rate.

Table 8. Age-sex distribution of Congolese migrants in the MAFE survey in Belgium, estimated age-sex distribution of Congolese migrants in Belgium and inflating factors.

	MAFE Congo-Belgium			Estimated age-sex structure			Inflating factors	
Age groups	**Males**	**Females**	**Total**	**Males**	**Females**	**Total**	**Males**	**Females**
25-34	31	26	57	2857	4668	7525	92	180
35-44	72	35	107	4255	5735	9990	59	164
45-54	35	38	73	3900	2831	6730	111	74
55 and over	16	26	42	1030	1725	2754	64	66
Total	154	125	279	12042	14958	27000		

UNITED KINGDOM – CONGOLESE MIGRANTS

Estimated Age-sex distribution of Congolese migrants in the UK

The age-sex distribution of Congolese migrants in the United-Kingdom is estimated using 3 data sources:

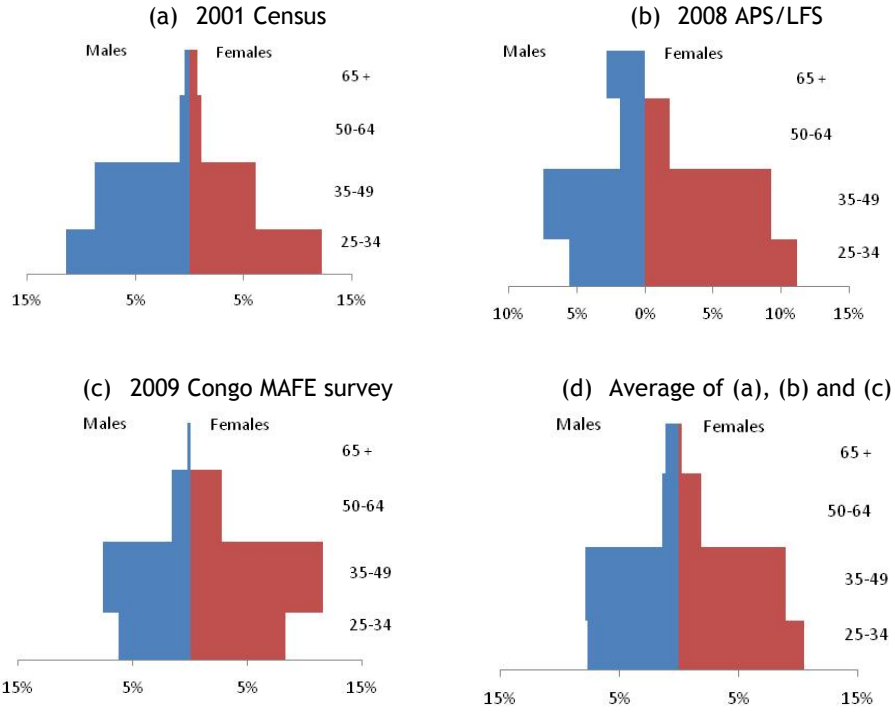
(a) The age-sex distribution of Congolese migrants in the 2001 Population census.

(b) The estimated-age sex distribution of Congolese migrants in the Annual Population Survey (APS)/Labour Force Survey (LFS) in 2008.

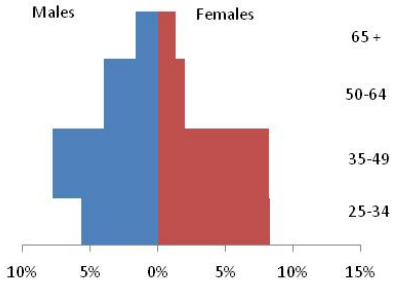
(c) The age-sex distribution of Congolese migrants in the UK identified in the MAFE survey conducted in Kinshasa (2009).

The relative age-sex distributions from these sources are represented on Figure 2(a) to Figure 2(c). The estimated age-sex distribution of the Congolese migrant population in the UK is estimated as the average of the three distributions (Figure 2(d)). Although this is a crude approach, the estimated age-sex distribution is realistic. The average age distribution is a little older than what was found in the 2001 census. This can reflect the fact that the Congolese immigration in the early 2000s was still very recent. With time, migrants get older, and the composition of immigrant flows may become less selective with regards to age, leading to an older age structure. The sex distribution indicates that females are more numerous than males (55% of females). This is also visible in the APS/LFS surveys (53% of females) and in the Congo MAFE surveys (62%), which are both more recent than the 2001 census.

Figure 2(a) to 2(e). Age-sex structures of Congolese migrants in the UK from various sources.



(e) 2009 UK-Congo MAFE survey



Post-stratification weights

Figure 2(e) and the first part of Table 3 show the age-sex distribution of the respondents in the Congo MAFE survey in the UK. Overall, males are overrepresented (50% compared to 46% in the average distribution), and older people are also (purposefully) overrepresented in the MAFE samples. Post-stratification weights are computed as the ratios of percentages in the average distribution to the percentages in the MAFE UK survey. Cells with large post-stratification weights (e.g. males and females below 35) indicate that migrants with these characteristics were underrepresented in the MAFE survey.

Table 9. Relative age-sex distribution of Congolese migrants in the MAFE survey in the UK, estimated age-sex distribution of Congolese migrants in the UK, and post-stratification weights.

	MAFE Congo-UK			Estimated relative age-sex structure			Post-stratification weights	
	Males	**Females**	**Total**	**Males**	**Females**	**Total**	**Males**	**Females**
25-34	11.3%	16.7%	28.0%	15.5%	21.1%	36.5%	1.36	1.27
35-49	23.3%	24.7%	48.0%	23.7%	26.9%	50.7%	1.02	1.09
50-64	12.0%	6.0%	18.0%	4.3%	5.7%	10.0%	0.36	0.94
65-74	3.3%	2.7%	6.0%	2.3%	0.5%	2.8%	0.70	0.18
Total	50.0%	50.0%	100%	45.8%	54.2%	100%		

Estimated Population of Congolese migrants aged 25-75 in the UK and inflation factors

The next step consists in estimating the population of Congolese migrants aged 25-75 in the UK at the time of the survey. According to the 2001 census, 8542 Congolese migrants were living in the UK, of which 5601 were aged 25 and over (ONS, 2009a). In 2008, the population of Congolese migrants was estimated at 15 000 according to the APS/FLS (ONS, 2009a), of which approximately 10 000 were aged 25-75. The MAFE survey data conducted in DR Congo can also be used to estimate the population of Congolese migrants in the UK. The basic approach is the following one. The MAFE survey in Congo indicates that about 162 migrants living in the UK were identified with the household survey, and 203 were living in Belgium. We consider that the relative weight of Congolese in Belgium and the UK is roughly correct. The size of the Congolese population aged 25-75 in Belgium in 2009 was estimated at 27000 (see above). Considering that the ratio of migrants in the UK to migrants in Belgium in the MAFE survey is roughly correct (0.80), this means that Congolese migrants in the UK aged 25-75 are about 21600 in 2009. This estimation is significantly

higher than the one provided by the other sources (2001 census and APS/LFS survey). However, a 2006 IOM report estimated migrants from RD Congo living in the UK to be between 20 000 and 40 000 (IOM, 2006). MAFE biographic surveys also indicate that undocumented migrants represent a signification proportion of Congolese migrants in the UK. In other words, 21600 Congolese migrants is a plausible estimate. The sampling rate of the MAFE UK-Congo survey is around 7/1000.

Table 10. Age-sex distribution of Congolese migrants in the MAFE survey in The UK, estimated age-sex distribution of Congolese migrants in the UK and inflating factors.

	MAFE Congo-UK			Estimated age-sex structure			Inflating factors	
	Males	**Females**	**Total**	**Males**	**Females**	**Total**	**Males**	**Females**
25-34	17	25	42	3306	4387	7693	194	175
35-49	35	37	72	5043	6190	11233	144	167
50-64	18	9	27	908	1165	2073	50	129
65-74	5	4	9	499	103	602	100	26
Total	77	72	149	9755	11845	21600		

Table 10 shows the age-sex structure in the MAFE sample, and the estimated numbers of Congolese migrants by age and sex in the UK. The inflating factors are obtained by computing the ratio of estimated population by age and sex to the sample size by age and sex in the MAFE UK sample. The inflating factors are also equal to the post-stratification weights divided by the sampling rate.

COMPUTATION OF WEIGHTS FOR THE GHANAIAN SAMPLES IN EUROPE

UNITED KINGDOM – GHANAIAN MIGRANTS

Estimated age-sex distribution of Ghanaian migrants in the UK

The age-sex distribution of Ghanaian migrants in the United-Kingdom is estimated using 3 data sources

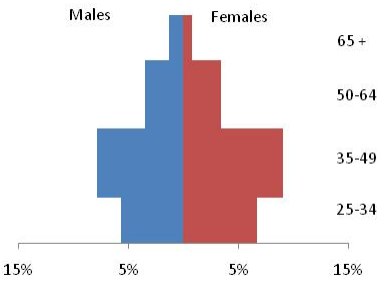
(a) The age-sex distribution of Ghanaian migrants in the 2001 Population census.

(b) The estimated-age sex distribution of Ghanaian migrants in the Annual Population Survey (APS)/Labour Force Survey (LFS) in 2008.

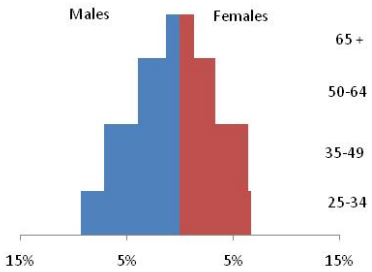
(c) The age –sex distribution of Ghanaian migrants in the UK identified in the MAFE survey conducted in Accra and Kumasi (2009).

Figure 3(a) to 3(e). Age-sex structures of Ghanaian migrants in the UK from various sources.

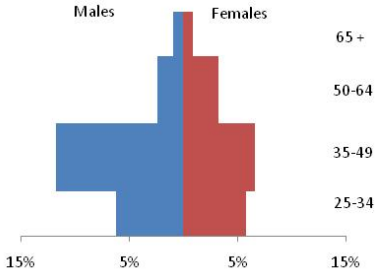
(a) 2001 Census



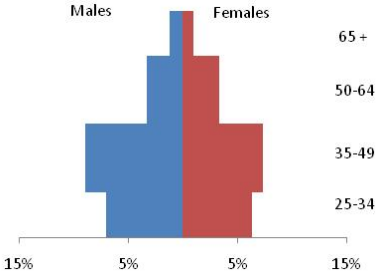
(b) 2008 APS/LFS



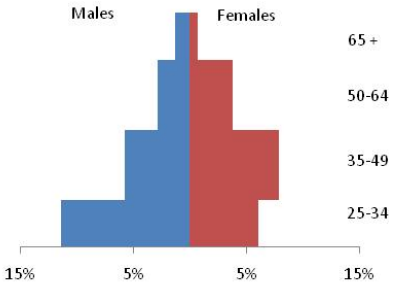
(c) 2009 Ghana MAFE survey



(d) Average of (a), (b) and (c)



(e) 2009 UK-Congo MAFE survey



The relative age-sex distributions from these sources are represented on Figure 3(a) to Figure 3(c). The estimated age-sex distribution of the Ghanaian migrant population in the UK is estimated as the average of the three distributions Figure 3(d). Although this is a crude approach, the estimated age-sex distribution seems realistic. The average age distribution is very similar to what was found in the 2001 census and the 2009 MAFE survey, which seem more reliable that the APS/LFS data in this regard. The sex distribution (53% of males) is close to the Ghana MAFE survey and the APS/LFS surveys, which are both more recent than the 2001 census.

Post-stratification weights

Figure 3(e) shows the age-sex distribution of the respondents in the Ghana MAFE survey in the United Kingdom. Young males (25-34) are overrepresented compared to their older counterparts and their female counterparts. Post-stratification weights are computed as the ratios of percentages in the average distribution to the percentages in the MAFE UK survey. Cells with large post-stratification weights (e.g. males between 35 and 64) indicate that migrants with these characteristics were underrepresented in the MAFE survey.

Table 11. Relative age-sex distribution of Ghana migrants in the MAFE survey in the UK, estimated age-sex distribution of Ghanaian migrants in the UK, and post-stratification weights.

	MAFE Ghana-UK			Estimated age-sex structure			Post-stratification weights	
	Males	**Females**	**Total**	**Males**	**Females**	**Total**	**Males**	**Females**
25-34	22.8%	12.1%	34.9%	14.2%	12.7%	26.9%	0.62	1.05
35-49	17.4%	23.5%	40.9%	26.8%	22.1%	48.9%	1.54	0.94
50-64	8.7%	11.4%	20.1%	9.9%	10.0%	19.9%	1.14	0.87
65-74	2.7%	1.3%	4.0%	2.4%	1.9%	4.3%	0.90	1.44
	51.7%	48.3%	100%	53.3%	46.7%	100%		

Estimated population of Ghanaian migrants aged 25-75 in the UK and inflation factors

The next step consists in estimating the population of Ghanaian migrants aged 25-75 in the UK at the time of the survey. Between the early 2000 and 2009, the Ghanaian population increased from around 56000 in the 2001 census (ONS, 2009a) to more than 93000 in July 2009 in the APS/LFS (ONS, 2009b). It seems that it then decreased, to reach around 85000 Ghanaian migrants in October 2009, (8 000 less than a year earlier) (ONS, 2009b). Although this estimate is not perfectly reliable (the 95% confidence interval ranges from 73000 to 97000 migrants), we consider this value (85 000) as our best estimate for the total population at the time of the survey. The percentage of migrants aged 25-75 or 25 and over was around 85% in the 2001 Census (83%) and in the APS/LFS in 2008 (86%). We use the value of 85% to estimate the percentage of Ghanaian migrants aged 25-75 at 72270. Using the age-sex structure of Table 11 and this population, we obtain the estimated age sex structure. The inflating factors are obtained by computing the ratio of estimated population by age and sex to the sample size by age and sex in the MAFE UK sample. The inflating factors are also equal to the post-stratification weights divided by the sampling rate. The sampling rate is about 2/1000.

Table 12. Age-sex distribution of Ghanaian migrants in the MAFE survey in the UK, estimated age-sex distribution of Ghanaian migrants in the UK and inflating factors.

	MAFE Congo-UK			Estimated age-sex structure			Inflating factors	
	Males	**Females**	**Total**	**Males**	**Females**	**Total**	**Males**	**Females**
25-34	34	18	52	10262	9178	19441	302	510
35-49	26	35	61	19368	15972	35340	745	456
50-64	13	17	30	7155	7227	14382	550	425
65-74	4	2	6	1734	1373	3108	434	687
Total	77	72	149	38520	33750	72270		

THE NETHERLANDS – GHANAIAN MIGRANTS

Estimated age-sex distribution of Ghanaian migrants in the Netherlands

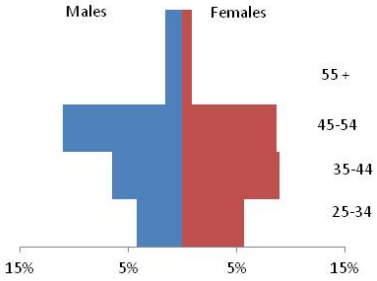
The age-sex distribution of Ghanaian migrants in the Netherlands is based on a single data source (Statistics Netherlands, CBS). Only 24 Ghanaian migrants (aged 25 and over) in the Netherlands were

identified in the MAFE survey conducted in Accra and Kumasi (2009), and could not be used to estimate an age-sex structure.

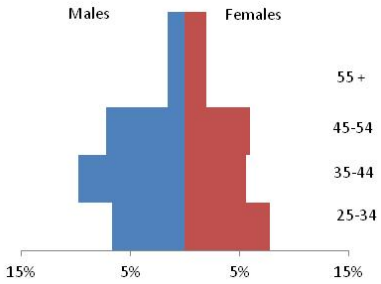
The relative age-sex distribution from this source is represented on Figure 3(a). The structure shows that young people are few compared to those aged 45-54. The sex distribution (50% of males and 50% of females) is balanced.

Figure 4(a) and 4(b). Age-sex structures of Ghanaian migrants in the Netherlands.

(a) CBS 2009



(b) 2009 Ghana MAFE survey - NDLS



Post-stratification weights

Figure 4b shows the age-sex distribution of the respondents in the Ghana MAFE survey in the Netherlands. Post-stratification weights are computed as the ratios of percentages in the average distribution to the percentages in the MAFE Netherlands survey. Cells with large post-stratification weights (e.g. males between 45 and 54) indicate that these people were underrepresented in the MAFE survey.

Table 13. Relative age-sex distribution of Ghana migrants in the MAFE survey in the Netherlands, estimated age-sex distribution of Ghanaian migrants in the Netherlands, and post-stratification weights.

	MAFE Ghana-Netherlands			Estimated relative age-sex structure			Post-stratification weights	
	Males	**Females**	**Total**	**Males**	**Females**	**Total**	**Males**	**Females**
25-34	13.3%	15.6%	28.9%	8.4%	11.3%	19.7%	0.63	0.73
35-44	19.6%	11.1%	30.7%	12.9%	17.9%	30.8%	0.66	1.61
45-54	14.4%	11.9%	26.3%	22.2%	17.4%	39.6%	1.53	1.47
55-74	6.3%	7.8%	14.1%	6.5%	3.3%	9.9%	1.04	0.43
Total	53.7%	46.3%	100.0%	50.1%	49.9%	100.0%		

Estimated population of Ghanaian migrants aged 25-75 in the Netherlands and inflating factors

We use the CBS data as the estimate of Ghanaian migrants in the Netherlands (CBS, 2009). It was estimated that 11 604 migrants from Ghana aged 25-75 were living in the Netherlands in 2009. This is much lower than in the UK. This lower number of migrants in the Netherlands is also visible in the household survey conducted in Ghana, where the number of migrants identified in the UK is about 10 times greater than the number of migrants in the Netherlands.

Table 14 shows the age-sex structure in the MAFE sample, and the estimated numbers of Ghanaian migrants by age and sex in the Netherlands. The inflating factors are obtained by computing the ratio of estimated population by age and sex to the sample size by age and sex in the MAFE UK sample. The inflating factors are also equal to the post-stratification weights divided by the sampling rate. The sampling rate is about 23/1000.

Table 14. Age-sex distribution of Ghana migrants in the MAFE survey in the Netherlands, estimated age-sex distribution of Ghanaian migrants in the Netherlands, and inflating factors.

	MAFE Ghana-Netherlands			Estimated age-sex structure			Inflating factors	
	Males	Females	Total	Males	Females	Total	Males	Females
25-34	36	42	78	976	1314	2290	27	31
35-44	53	30	83	1502	2077	3579	28	69
45-54	39	32	71	2572	2018	4590	66	63
55-74	17	21	38	759	386	1145	45	18
Total	145	125	270	5809	5795	11604		

COMPUTATION OF WEIGHTS FOR THE SENEGALESE SAMPLES IN EUROPE

Senegalese migrants in France and Italy were selected using quota sampling with quotas set by age groups and sex, while the sample in Spain was a probabilistic sample from the population register, augmented by Senegalese respondents identified through contacts in the household survey in Senegal. Given this mix, weights for Spain were also computed by poststratification and not by computing sampling probabilities.

The procedure for constructing post-stratification weights in the case of MAFE-Senegal consisted of the following steps:

- Computation of target population estimates
- Computation of initial inflation weights
- Computation of poststratification weight.

The following sections describe the procedure in each of the three countries.

FRANCE – SENEGALESE MIGRANTS

Computation of target population estimates

Estimates of the distribution of the target population in terms of age and sex characteristics were computed based on auxiliary data. For the case of France, data on individuals born in Senegal were available from the Renovated Population census (RPP; INSEE, 2004-2007). However, the available data source did not coincide exactly with the target population criteria defined for the MAFE surveys. Target population estimates had to be computed to introduce the age restriction to individuals of age 25 or older and the geographical restriction to three geographical areas: Ile de France, around Paris; Rhône-Alpes, around Lyon; and Provence-Alpes-Côte d'Azur (PACA), around Marseille.

For instance, Table 15 presents the distribution Senegalese migrants in the census and the MAFE data by their geographical location and sex. However, the RRP data contain **all ages**, not only the age groups considered by MAFE data. Data are therefore not directly comparable.

Table 15. Distribution of RRP and MAFE data by geographical area and sex, absolute and relative frequencies

Geographical areas	Absolute frequencies					
	Females		Males		Total	
	RRP	MAFE	RRP	MAFE	RRP	MAFE
Ile de France	22555	71	25525	84	48080	155
PACA	4427	13	4689	14	9116	27
Rhône Alpes	2799	8	3417	11	6216	19
Total	29781	108	33631	92	63412	200
Geographical areas	Relative frequencies					
	Females		Males		Total	
	RRP	MAFE	RRP	MAFE	RRP	MAFE
Ile de France	76%	66%	76%	90%	76%	77%
PACA	15%	12%	14%	15%	14%	14%
Rhône Alpes	9%	7%	10%	12%	10%	10%
Total	100%	100%	100%	100%	100%	100%

The next step is therefore to obtain information about the share in terms of ages considered in MAFE and to adjust population totals correspondingly. The share of the population aged 25 years or older is 83% in Ile-de-France and PACA and slightly lower in Rhône-Alpes. The estimated reference population in the three regions and satisfying the age criteria is 52559.

Computation of initial weights

For France, a single initial weight is computed as the estimated target population divided by the number of individuals interviewed in MAFE-France (52559/201=261.5).

Computation of post-stratification weights

In a third step, post-stratification weights were constructed in order to adjust weighted frequencies to population frequencies in auxiliary data with regard to sex and age in three categories (25-34 years, 35-44 years, and 45-75 years). Since no full matrix was available in the case of France, a calibration method called raking ratio was applied, which consists of an iterative adjustment of weights based on the marginal population totals according to sex and three age groups (Table 16).

Table 16. Auxiliary data used in raking ratio poststratification - France

Males	Females	25-34 years	35-44 years	45 years +	Total
27 869	24 690	13 276	14 477	24 806	52 559

Table 17 presents summary statistics of the inflation weight variable obtained for France.

Table 17. Summary statistics inflation weight France

N	Sum	Mean	Min	Max
201	52559	261.5	188.28	378.4

ITALY – SENEGALESE MIGRANTS

Computation of target population estimates

Estimates of the distribution of the target population in terms of age and sex characteristics were computed based on auxiliary data. For the case of Italy, two auxiliary data sources were used. Firstly, the data on residence permits for Senegalese citizens (ISTAT, as of 01/01/2006), and secondly, estimates of the number of Senegalese irregular migrants by region from the ISMU (Iniziative e Studi sulla multietnicità) surveys directed by Gian Carlo Blangiardo (2007).

As in France, the available data source did not coincide exactly with the target population criteria defined for the MAFE surveys. Target population estimates had to be computed to introduce the age restriction to individuals of age 25 or older and the geographical restriction to four geographical areas: Lombardia, Emilia-Romagna, Toscana, and Campania. Two additional adjustments were necessary in the Italian case. The target population estimate had to consider individuals who had acquired the host country citizenship, as permit data only included information by citizenship and not by country of birth. Also, an estimate of the number of irregular migrants had to be computed.

The next steps were hence the following. Firstly, the gross data on regular migrants by sex and in two age groups (Table 18), which are for the whole of Italy, were adjusted by the share in the four regions covered in MAFE (65%) by multiplying the cells by 0.65.

Table 18. Gross data on regular Senegalese citizens by sex and age (ISTAT, 2006)

	25 to 39 years	40 years+
Females	4245	1072
Males	20238	18923

This step implies the assumption that the Senegalese regular population is distributed equally across the regions by age and sex. In a next step, estimates are adjusted to account for the share of naturalized Senegalese. Since no external data source was available, the share had to be computed based on the MAFE data themselves. The share is estimated as ranging between 2% for males in the 25-39 age group to 14% in the 40 years plus age group. The following step adjusts for the share of irregular migrants. The available data provides estimates by region, but not by age or sex. The share of irregular migrants is thus computed for the four regions covered in MAFE (average of 20.95) and the percentage is applied to the previously estimated age-sex distribution to obtain estimates of irregular migrants. Once again, the (strong) assumption is that the age-sex distribution of irregular migrants is comparable to the age-sex distribution of regular migrants. The target population estimates (by age/sex) are then computed as the sum of the estimated number of regular citizens in the four regions, the estimated number of naturalized Senegalese immigrants, and the estimated number of irregular migrants.

As one can see in Table 19, women were on purpose greatly overrepresented in the MAFE data.

Table 19. Distribution of estimated target population and MAFE data - Italy

Gender	Age groups	Estimated target population	MAFE
Females	25-39 years	9.7%	25.4%
	40 years +	2.6%	13.2%
Males	25-39 years	45.0%	31.2%
	40 years +	42.7%	30.2%
		100%	100%

Computation of initial weight

Initial inflation weight has been constructed as the ratio the estimated population totals and the number of individuals interviewed in the survey, accounting for the sex and age-distribution (in 2 categories).

	25-39 years	40 years +
Females	59.77	41.68
Males	246.36	249.25

Computation of post-stratification weights

In a third step, post-stratification weights were constructed in order to adjust weighted frequencies to population frequencies in auxiliary data with regard to sex and age in three categories (25-34 years, 35-44 years, and 45-75 years).

Table 20. Estimated target population used in poststratification - Italy

Males			Females			Total
25-34 years	35-44 years	45 years +	25-34 years	35-44 years	45 years +	
7 994	14 846	7 901	2 370	1 552	384	35 047

Table 21 presents summary statistics of the inflation weight variable obtained for Italy.

Table 21. Summary statistics inflation weight Italy

N	Sum	Mean	Min	Max
205	**35047**	**172.6**	**25.3**	**284.5**

SPAIN – SENEGALESE MIGRANTS

Computation of target population estimates

Estimates of the distribution of the target population in terms of age and sex characteristics were computed based on auxiliary data. For the case of Spain, the population register (padrón, INE; 01/01/2007) provided auxiliary data for Senegalese citizens in Spain, both regular and irregular. The available data source did not coincide exactly with the target population criteria defined for the MAFE surveys. Data by age and sex are relatively detailed in the pardon. However, as in the case of Italy, an adjustment for individuals who had acquired the host country citizenship was introduced.

As one can see in Table 19, (elderly) women were greatly overrepresented in the MAFE survey.

Table 22. Distribution of estimated target population and MAFE data - Spain

		Estimated target population	MAFE
Females	**25-39 years**	**10.0%**	**32.0%**
	40 years +	**4.5%**	**28.0%**
Males	**25-39 years**	**59.9%**	**22.5%**
	40 years +	**25.6%**	**26.5%**
		100%	**100%**

Computation of initial weight

Initial inflation weight has been constructed as the inverse of individuals interviewed in the survey and the estimated population totals, accounting for the sex and age-distribution (in 2 categories).

	25-39 years	40 years +
Females	61.31	24.72
Males	384.17	144.98

Computation of post-stratification weights

In a third step, post-stratification weights were constructed in order to adjust weighted frequencies to population frequencies in auxiliary data with regard to sex and age in three categories (25-34 years, 35-44 years, and 45-75 years).

Table 23. Estimated target population used in poststratification - Spain

Males			Females			Total
25-34 years	35-44 years	45 years +	25-34 years	35-44 years	45 years +	
12 502	8 494	4 056	2 138	1 495	804	29 489

Table 24 presents summary statistics of the inflation weight variable obtained for Italy.

Table 24. Summary statistics inflation weight Spain

N	Sum	Mean	Min	Max
200	29489	147.4	24.0	568.3

The large max/min ratio shows the extent to which female/elderly migrants had been overrepresented in the MAFE survey data with respect to the population, characterized by relatively young, male Senegalese immigration.

COMPUTATION OF NORMALIZED WEIGHTS IN THE MAFE DATA

In the MAFE data, all survey weights have been rescaled (normalized) so that the sum of weights corresponds to the sample sizes of households and individuals respectively while the mean of the weight variables equals one. It is essential that the sum of weights is equal to the sample size when statistical tests are performed; otherwise, standard errors would be underestimated. When several datasets are pooled together, using normalized weights in each data set separately is not appropriate, because doing so assumes that the sampling fraction is similar in each data set. It is thus necessary to computed normalized weights for each combination of datasets.

NORMALIZED WEIGHTS IN HOUSEHOLD SURVEYS

As explained before, weights are computed as the inverse of the sampling probabilities of households, and are then trimmed and adjusted so that the sum of weights in the household surveys is equal to the total number of households in the city (N)[16]. The adjusted weights are noted w_h^{**}.

[16] In practice, it is easier to ensure that the sum of weighted individuals in the household surveys is equal to the total population of the city, because the size of the population can be estimated more easily (United Nations, 2009). Considering that the size of household is correct in the MAFE survey, the two methods are equivalent.

These are also called inflation factors. The inflation factor can be interpreted as measuring “how many households” each household in the sample represents.

$$\sum_{h=1}^{n} w_h^{**} = N$$

Normalizing weights in this context consists in transforming w_h^{**} so that their sum is equal to the sample size of households (n). If p_h is the normalized weight, this condition is written:

$$\sum_{h=1}^{n} p_h = n$$

p_h is equal to the inflation factor multiplied by a constant c. The constant is simply equal to the ratio of the sample size of households (n) to the number of households (N).

$$\sum_{h=1}^{n} c.w_h^{**} = n$$

$$c.\sum_{h=1}^{n} w_h^{**} = n$$

$$c.N = n$$

$$c = n/N$$

The normalized weight is thus simply equal to the inflation factor divided by the number of households in the population, multiplied by the sample size. It can also be c omputed in straightforward way, because c is also the inverse of the mean of the inflation factor: dividing each inflation factor by the mean inflation factor gives a normalized weight.

$$c = \frac{n}{\sum_{h=1}^{n} w_h^{**}}$$

$$p_h = c.w_h^{**} = \frac{w_h^{**}}{\sum_{h=1}^{n} w_h^{**}/n}$$

NORMALIZED WEIGHTS IN BIOGRAPHIC SURVEYS

The principle for computing normalized weights in the biographic survey is similar to the one used in the household survey. Inflation factors in the biographic surveys are noted w_b^{**}. The inflation factor can be i nterpreted as measuring “how many individuals” each individual in the sample represents.

The sum of inflation factors in the biographic surveys is equal to the size of the population covered by the survey. In African cities, the sum of weights in the biographic survey is equal to the population of the city aged 25-75. For instance, the sum of weights in Kinshasa is equal to 2 552 870 individuals. In Europe, the sum of weights is equal to the estimated size of the migrant population

(defined with the MAFE eligibility criteria). For instance, the sum of weights among Ghanaians in the UK is equal to 72 270.

$$\sum_{b=1}^{n} w_b^{**} = N$$

When two or more biographic surveys are pooled together, the sum of the inflation factors is equal to sum of the sizes of the populations corresponding to each survey.

For instance, the biographic surveys among Ghanaians in the UK and in the Netherlands can be pooled together. The sum of inflation factors is equal to 72 270 (Ghanaians in the UK) + 11 6 04 (Ghanaians in the Netherlands).

$$\sum_{b=1}^{n1+n2} w_b^{**} = N1 + N2$$

If the biographic surveys among Congolese migrants in Belgium and in the UK, and the biographic survey in Kinshasa are pooled together, we have

$$\sum_{b=1}^{n1+n2+n3} w_b^{**} = N1 + N2 + N3$$

Where N1 is the population size in Kinshasa (2 552 870), N2 the size of the Congolese population in Belgium (27 000), and N3 the size of the Congolese population in the UK (21 600).

We briefly describe the computation of normalized weights in this context (2 European and 1 African sample), but the approach is very general.

Normalizing weights consists in transforming w_b^{**} so that their sum is equal to the sample size of individuals (for instance n1+n2+n3). If p_b is the normalized weight, this condition is written:

$$\sum_{b=1}^{n1+n2+n3} p_b = n1 + n2 + n3$$

p_b is equal to the inflation factor multiplied by a constant c. The constant is equal to the ratio of the sample size of individuals (n1+n2+n3) to the number of individuals (N1+N2+N3).

$$\sum_{b=1}^{n1+n2+n3} c.w_b^{**} = n1 + n2 + n3$$

$$c. \sum_{b=1}^{n1+n2+n3} w_b^{**} = n1 + n2 + n3$$

$$c.(N1 + N2 + N3) = (n1 + n2 + n3)$$

$$c = (n1 + n2 + n3)/(N1 + N2 + N3)$$

The normalized weight is thus equal to the inflation factor divided by the total number of individuals in the pooled populations, multiplied by the total sample size. It can also be computed

in straightforward way, because c is also the inverse of the mean of the inflation factor: dividing each inflation factor by the mean inflation factor gives a normalized weight.

$$c = \frac{n1 + n2 + n3}{\sum_{h=1}^{n1+n2+n3} w_b^{**}}$$

$$p_h = c.w_b^{**} = \frac{w_b^{**}}{\sum_{h=1}^{n1+n2+n3} w_h^{**} / (n1 + n2 + n3)}$$

TYPES OF NORMALIZED WEIGHTS IN THE MAFE DATA BIOGRAPHIC DATA SETS

Several types of normalized weights are available in the MAFE biographic data sets.

- weight_**all** is a normalized weight variable computed after pooling all the biographic data sets for one specific origin country. For instance, Ghanaians in the UK, Ghanaians in the Netherlands, and Ghanaians in Ghana are pooled together. Their normalized weights are computed like in the example described in the previous section (3 samples pooled together).
- weight_**eur is** a normalized weight variable computed after pooling the biographic data sets for one specific origin country only in the destination countries. For instance, Ghanaians in the UK and Ghanaians in the Netherlands are pooled together.
- weight_**ctry is** a normalized weight variable computed only in one country.

TO WEIGHT OR NOT TO WEIGHT? WHAT DOES THE LITERATURE SUGGEST?

This section provides a brief summary of the literature on the justification of using or not using weights in descriptive and multivariate analyses.

DESCRIPTIVE ANALYSIS

The common recommendation is to account for the sampling design in descriptive summary statistics, such as means, medians, frequency distributions of single variables or cross-tabulations, in order to correct for unequal selection probabilities of units in the sample and varying response-rates over sub-populations (Pfefferman, 1993; Chromy and Abeyasekera, 2005). Kish (1965) points out the exception of equal probability samples, which may be achieved for the final sampling units even in multistage designs. Most often, however, weights should be a pplied in the context of descriptive analysis in order to infer from sample statistics to finite population parameters, i.e. of parameters describing the population from which the sample was drawn.

REGRESSION ANALYSIS

The literature remains divided with regard to the use of survey weights in regression analysis (see, for instance, DuMouchel and Duncan, 1983; Winship and Radbill, 1994; Deaton, 1997; Kalton, 2002; Little, 2004; Lee and Forthofer, 2006), and different viewpoints are in line with either a design-based approach or a model-based/superpopulation approach.

On the one hand, the *design-based or frequentist perspective* assumes that "sample data are observations sampled from a finite population using a particular sample selection design", which indicates the probability of selection of each potential sample (Lee and Forthofer, 2006). Inference is made to finite population quantities. Regression is regarded as descriptive and provides a device

to summarise characteristics of the population. According to this view, survey weights and survey design should be accounted for in all types of analysis.

On the other hand, the *model-based view* stipulates that observations in the finite population are seen as realisations of a random variable generated from a model, describing, for instance, an economic process (Deaton, 1997; Pfefferman, 1993). Models are thus used to draw inference to so called infinite superpopulations that are more general than the finite population from which the sample was drawn. Following some probability distribution, the model allows for prediction of unobserved values based on observed values in the sample. Inference is a prediction problem, and based on the joint distribution of the survey outcomes *Y* and the set of variables *I* indicating whether a unit is included in the sample or not. Under this approach, use of sampling weights is not necessary, under the condition that the observations follow the model and as long as the selection probability depends on the dependent variable of the model only through the independent variables included. This implies that the sampling design is ignorable/non-informative for the analysis at hand, meaning that selection probabilities are uncorrelated with variables of interest (when conditioned on explanatory variables).

Table X: Summary design-based and model-based approaches to survey analysis

	Design-based	Model-based
Advantages	- Automatically takes into account features of the survey design (no need to include them in the model specification) - Provides reliable inferences in large samples - No need for strong modelling assumptions	- More efficient if correctly specified - Based on substantive theory and previous empirical investigations and can hence account for the case of population heterogeneity
Limitations	- Asymptotic, hence limited guidance for small sample adjustments - No prescription for choice of estimator - Lacks theory for optimal estimation - Estimates tend to be less efficient - Weighted analysis heavily influenced by observations carrying large weights - Weights are no "simple solution" to model misspecification arising from other aspects, one needs to anyhow think about how to correctly model behaviour - Does not account for dynamic nature of populations (changes in population between date of survey and date of inference)	- May produce biased estimates and standard errors if model is not correctly specified with regard to the inclusion of relevant design variables; non-response and response errors, and survey design is related to the survey outcomes analysed

Sources: Hoem (1989), Pfefferman (1993), Lee and Forthofer (2006), Deaton (1997), Kalton (2002)

No generic recommendation can be made as to whether to account for the sampling design in regression analysis. Decisions have to be made by the analysts depending on the dependent variable to be explained, information about independent variables that can be included in the estimation, and the relation to variables used in the sampling design. Also, opinions are often divided based on disciplinary lines, with statisticians favouring a design-based approach and econometricians favouring a model-based approach.

However, even if a model-based approach is adopted, comparing weighted and unweighted estimates can represent a useful exercise, as one may be able to identify variables or interactions that should be included in the estimation to avoid model misspecification.[17]

[17] Moreover, several authors have proposed statistics to test for differences in point estimates between weighted and unweighted estimates (see DuMouchel and Duncan, 1983; Fuller, 1984). However, testing for ignorability of the design is not straightforward.

EVENT-HISTORY ANALYSIS

Biographic data collected retrospectively in the MAFE surveys are particularly suited for analyses of life-history analysis and the hazards of events, such as migration, birth of a child, marriage, or investment using event-history analysis techniques.[18] Several experts in the area of life-history analysis are favouring the model-based approach (Hoem, 1985; Courgeau and Lelièvre, 1992). However, the authors also acknowledge that analysis based on retrospective data may be more likely to be subject to informative sampling than cross-section data or panel data (Hoem, 1985; Courgeau and Lelièvre, 1992; Neuhaus and Jewell, 1990). Informative sampling implies that the probability of selection into the sample depends on previous behaviour (such as having migrated), and the outcome of the behaviour may be subject to analysis. In this case, sampling design features are related to the outcome and need to be accounted for to guard against selection biases. As in the general case of regression analysis, whether one needs to account for sampling design depends thus on the research question and modelling approach. A particular case is discrete-time event history analysis with repeated events and/or unobserved heterogeneity, which can be estimated as binary dependent variable model with a random intercept. While the literature on using sampling weights in multilevel models is growing (e.g. Grilli and Pratesi, 2004; Zaccarin and Donati, 2008; Carle, 2009), the properties of estimates remain less explored and statistical software packages do not always allow for weighting and correction of the variance.

ANALYSES ON POOLED DATASETS (SEVERAL COUNTRIES)

When pooling samples for different populations, the relative weight to be given to each sample is dependent on substantive considerations. In many applications, country-level weights are adjusted by to be proportional to its population size. For instance, samples in countries conducting European Social Surveys are often similarly in size, despite the large differences in population sizes (Kish, 1994; Skinner and Mason, 2012). Scaling will give again more weight to countries with large populations and reduce the weight of countries with small populations, and statistics will be produced for the "average EU citizen". However, depending on the research question, other scaling procedures can be applied to the weights. Skinner and Mason (2012), for instance, explore methods to modify country-level design weights for cross-national pooled analysis. In general, model-based analysis involving the inclusion of design variables as explanatory variables is particularly challenging in analyses on pooled data, as sampling designs may differ across countries. Moreover, as Thompson (2008) remarks: *"an analysis which pools data across countries should be adopted with caution. For such an analysis to be appropriate, the model structure (the regression equation and its variables) should be correct for all countries, and the assumption of common parameters should be supported by theory and observation".*

Given that not many surveys sample populations across borders, the discussion of weights in pooled analyses refers mainly to cross-national surveys, such as the European Social Survey. One exception is the Mexican Migration Project, which surveyed Mexican households in communities in Mexico as well as migrants originated from those communities in the United States (MMP, http://mmp.opr.princeton.edu/databases/studydesign-en.aspx). Migrants were predominantly selected based on referrals from the origin. The weights for the US sample are computed based on information on number of children who settled in the United States versus those who left the parental home but stayed in Mexico. Applying "community-specific" weights in pooled analysis is hence supposed to produce data that are representative of the population of all communities, whether they live in Mexico or in the US.

[18] Event-history analysis can also be performed on prospectively collected longitudinal data (panel surveys). Weights in panel surveys usually include both cross-sectional design weights as well as longitudinal weights that account for the change in population structure over panel waves.

USING MAFE WEIGHT (AND DESIGN) VARIABLES

Given the different stances researchers can take over weighting, this section intends to provide suggestions rather than prescriptions.

ANALYSIS EITHER OF THE SENEGALESE, THE GHANAIAN, OR THE CONGOLESE SAMPLE

Each of the African samples is a probability sample representative at the region level. Depending on the research question and method, either a model-based or a design-based approach can be chosen when analyzing household or biographic data. In the latter case, survey weight variables and design variables summarized in Tables 3 and 4 can be applied to the analysis. For descriptive analyses, it is recommended to use the weight variables in order to control for the complex sampling design.

ANALYSIS POOLING DATA FROM SEVERAL SAMPLES

European samples of migrants will be in most cases too small for separate analysis and analysts may therefore consider to pool data on one migrant flow in various European countries. For example, Senegalese in France, Spain or Italy may be analyzed as "Senegalese migrant population in main European destination countries". Weights for pooled analyses are normalized/ scaled to reflect the population size of a migrant group in each destination country.

As explained above, European samples are constructed using quota sampling rather than a probabilistic method. Weights are hence poststratification weights on a limited number of population characteristics. However, we generally recommend the use of weights in descriptive analyses to account for the fact that elderly and women were overrepresented in most samples. We would like to emphasize limitations of the use of weights in analyzing European samples:

- The poststratification is valid under the strong assumption that the quota sample is similar to a stratified random sample. However, given the diversity of sampling methods, it is likely that selection biases exist, that are not accounted for by applying the poststratification weights.
- Since available auxiliary data did not always provide full matrices or at least margins for calibration to the eligibility criteria and quotas established in the MAFE surveys, the poststratification is performed on estimates. Estimation of target population totals often relied on relatively strong assumptions.

Depending on the research question, analysts may want to pool data for one flow (Senegal, Ghana, or DR Congo) across European and African samples. Researchers should be aware of several implications of this type of data use. Firstly, analysts should take into consideration that the pooling probability and non-probability samples rests on the assumption of ignorability of the unknown sampling mechanism in the non-probability samples, i.e. the factors that determine a population member's presence or absence in the sample are all uncorrelated with the variables of interest in a study, or they are fully accounted for by the use of quotas and poststratification weights (Yaeger et al., 2011). Secondly, researchers should be aware that pooling and application of normalized weights (weight_**all**) for the flow in question implies that inference is made to a transnational Senegalese population, taking account of population size. More weight is given to observations sampled in Africa than in Europe, where the ratio between sample size and African migrant populations is larger than between samples and population in the targeted African regions.

CORRECTING THE VARIANCE ESTIMATES

In Stata, finite population corrections (provided in the form of the sampling fraction in the MAFE database) account for the fact that sampling was done without replacement, which engenders a reduction in variance. However, the smaller the sampling fraction is, the smaller the reduction in variance. It is common to drop the finite population correction information when specifying the

sampling design. In this case, design variables at levels below the primary sampling units are ignored in the variance estimation.

BOX 1: Survey commands in Stata

If one is able to identify primary sampling units and strata variables in addition to the weight variable, one should use the survey commands: Svyset, svy estimation, svy postestimation. The data is identified as survey data and the sampling design is specified with svyset, and all survey commands (starting with svy:) "remember" the information about the survey design set at the beginning and compute point estimates AND standard error estimates accordingly. This approach is preferable to using commands which allow for the inclusion of the weight variable within each command [pweight=variable], since standard errors are not corrected when using pweight (a step towards the svy option would be to cluster standard errors by the cluster variable, subject to a minimum number of clusters). However, although svy commands have been further developed in recent Stata versions, there may be some estimation commands without the svy option, in which the pweight specification needs to be used.

When performing analysis on subgroups in a dataset that has been declared to a survey data (e.g. only women), one should keep the entire data set and assign weights of zero to observations one wants to exclude from the analysis (see 'subpop()' and 'over()' options in STATA).

While software packages differ with regard to the range of statistics and estimations supported, other providers such as SAS (proc survey commands); the R survey package or SPSS (complex samples modules) also allow for analysis with survey weights and standard error corrections.

REFERENCES

Beauchemin, C. (2012). Migrations between Africa and Europe: Rationale for a Survey Design, *MAFE Methodological Note* 5, 1-45.

Blangiardo, Gian Carlo (2007): Foreigner's Presence in Italy. Quantitative Evaluations and Comments. ISMU - The 12th Italian Report on Migration 2006. V. Cesareo, *Polimetrica*: 41-58.

CBS (2009), *Bevolking; leeftijd, herkomstgroepering, geslacht en regio, 1 januari, 2009,* CBS, Den Haag.

Chromy, J.R. and Abeyasekera, S. (2005). Statistical analysis of survey data. Chapter XIX in United Nations DESA (Ed.) *Household Sample Surveys in Developing and Transition Countries*, Studies in Methods Series F No. 96.

Deaton A. (1997). *The Analysis of Household Surveys*, World Bank; John Hopkins University Press.

Hoem, J. (1989). The issue of weights in panel surveys of individual behavior. Pp. 539-565 in Daniel Kasprzyk, Greg Duncan, Graham Kalton and M. P. Singh (eds.): *Panel Surveys*. New York: Wiley.

Hoem, J. (1985) Weighting, misclassification, and other issues in the analysis of survey samples of life histories; pp. 249-293 in Heckman and Singer (eds.): *Longitudinal analysis of labor market data*. Econometric Society Monographs.

Kalton, G. (2002). Models in the Practice of Survey Sampling (Revisited). *Journal of Official Statistics* 18(2), 129-154.

Kish, L. (1965). *Survey Sampling*. John Wiley & Sons, Inc., New York, London.

Lee, E.S. and Forthofer, R.N. (2006). *Analyzing complex survey data* (2nd edition). SAGE Publications 71 (Series Quantitative Applications in the Social Sciences).

Little, R. (2004). To model or not to model? Competing modes of inference for finite population sampling. *Journal of the American Statistical Association* 99(466), 546-557.

Neuhaus, J.M. and Jewell, P. (1990). The Effect of Retrospective Sampling on Binary Regression Models for Clustered Data, *Biometrics* 46(4), 977-990.

ONS (2009a), *Table C1044b Country of Birth (Ghana; Democratic Republic of Congo) by Usual address one year before Census by age (5 groups) by sex and Highest Educational Attainment (based on Highest Level of Qualification)*, created by ONS, London.

ONS(2009b), *Annual Population Survey (APS)/Labour Force Survey (LFS), Estimated population resident in the United Kingdom, by foreign country of birth*, ONS, London.

Pfeffermann, D. (1993). The role of sampling weights when modelling survey data. *International Statistical Review* 61(2), 317-337.

Razafindratsima, N., Legleye, S., and Beauchemin, C. (2011). Biais de non-réponse dans l'enquête Migrations entre l'Afrique et l'Europe (MAFE-Senegal), *MAFE Methodological Note* 4, 1-5.

Schoumaker, B., and Diagne, A. (2010). Migrations between Africa and Europe: Data Collection Report, *MAFE Methodological Note* 2, 1-28.

Winship, C. and Radbill, L. (1994). Sampling weights and regression analysis. *Sociological Methods and Research* 23, 230-257.

United Nations (2009), World Urbanization Prospects: the 2009 revision, United Nations, New York.

Yaeger, D.S., Krosnick, J.A., Chang, L., Javitz, H.S., Levendusky, M.S., Simpser, A., Wang, R. (2011). Comparing the accuracy of RDD telephone surveys and internet surveys conducted with probability and non-probability samples, *Public Opinion Quarterly* 75, 709-747.

ANNEX 1: SELECTION GRID OF RESPONDENTS FOR THE BIOGRAPHIC SURVEY

Household questionnaire number |__|__|__|__|

SELECTION OF ELIGIBLE RESPONDENTS FOR THE INDIVIDUAL SURVEY

1. Selection of return migrants currently member of the household

1. In Table 2, write down the given name, sex, age and line number of Module A of all the return migrants who are currently members of the household, aged 25-75, who were born in Ghana and have ever had Ghanaian citizenship.

2. Selection of partners/spouses of migrants

1. In Table 2, write down the given name, sex, age and line number of Module A of all the partners/spouses of migrants, who are currently members of the household, aged 25-75, who were born in Ghana and have ever had Ghanaian citizenship.

3. Random selection of another member of the household

1. In Table 1 below, write down the given name, sex, age and line number of Module A of all the other current members of the household aged 25-75, who were born in Ghana and have ever had Ghanaian citizenship (excluding return migrants and partners/spouses of migrants).
2. Take the last digit of the household questionnaire number (on the cover sheet and on the top of this page), and circle the corresponding digit in the Table 1 (line of digits from 1 to 0). Select the column corresponding to that digit, and select the line of the last person recorded in the table. Circle the number located at the intersection of the selected column and the selected line.
3. This is the number of the randomly selected person for the individual survey. For example, if the number is 3, select the person on the third line (rank equal to 3).
4. Write here: A) The rank of the selected person : NRANK : [__|__]
   B) The line number in module A of the selected person : NLINEAL : [__|__]
5. Write down the given name, sex, age and line number of Module A of the selected person on the last line of Table 2.

TABLE 1 : RANDOM SELECTION OF A HOUSEHOLD MEMBER

Rank	Given name	Sex 1. M 2. F	Age	Line number in Module A	Last digit of household questionnaire number									
					1	2	3	4	5	6	7	8	9	0
1					1	1	1	1	1	1	1	1	1	1
2					2	1	2	1	2	1	2	1	2	1
3					1	2	3	1	2	3	1	2	3	1
4					1	2	3	4	1	2	3	4	1	2
5					4	5	1	2	3	4	5	1	2	3
6					4	5	6	1	2	3	4	5	6	1
7					3	4	5	6	7	1	2	3	4	5
8					3	4	5	6	7	8	1	3	4	5
9					2	3	4	5	6	7	8	9	1	2
10					1	2	3	4	5	6	7	8	9	10

TABLE 2 : LIST OF HOUSEHOLD MEMBERS TO INTERVIEW

TYPE	GIVEN NAME	SEX	AGE	LINE NUMBER IN MODULE A
RETURN MIGRANT				
PARTNER/SPOUSE OF RETURN MIGRANT				
OTHER RANDOMLY SELECTED INDIVIDUAL				

ANNEX 2: STATA SYNTAX FOR SAMPLING HOUSEHOLDS IN ACCRA

```
**************************
* July 2009 – B. Schoumaker****
**************************


**************************
**SETTING SEED and PATH****
**************************
clear
set seed 100

cd "F:\0.MAFE\Sampling\Sampling Ghana\"

foreach y of numlist 1/60 {

clear
***********************************
**SET SAMPLE SIZE IN PSU************
***********************************

* ssizeC is the number of households to select in each category (it will adapated automatically if number of households lower than ssizeC)

local ssizeC=8
local ssizeT=3*`ssizeC'

***********************************
**READING SPREADSHEET***************
***********************************

insheet using Ghana_Accra.txt

keep if eanum==`y'

*** creation of  id
gene id=_n

***********************************
*** creation of migration status ***
***********************************

*** 0: no migration or other
*** 1: migrants abroad
*** 2: return migrants in the household

gene migstat=0
replace migstat=1 if partn=="Y"
replace migstat=2 if retm=="Y"

sort migstat id

**************************************************
** number of individuals in each category (stratum)***
```

```
**************************************************

by migstat, sort: egen numcat=count(migstat)
sort numcat migstat id


**************************************************************************
*** creation of indicators of strata (1=smallest number of individuals)**
**************************************************************************

egen strata=group(numcat migstat)

** number of strata

egen stratmax=max(strata)

*************************************
***SELECTION IN FIRST STRATUM**********
*************************************

*** NUMBER OF HOUSEHOLDS IN STRATUM 1

count if strata==1
local ncat1=r(N)

**** IF 2 strata  or more, sample in first stratum is either ncat1 or 8 (the smallest)

local ssize1=min(`ncat1', `ssizeC')
sample `ssize1', count, if strata==1 & stratmax>1

**** IF 1 stratum, the 24 households are selected in the first stratum

local ssize1b=min(`ncat1', `ssizeT')
sample `ssize1b', count, if strata==1 & stratmax==1

*************************************
***SELECTION IN SECOND STRATUM**********
*************************************

*** NUMBER OF HOUSEHOLDS IN STRATUM 2

count if strata==2
local ncat2=r(N)

**** if 3 strata, the sample in the second strata is equal to the lowest value of (ncat2, or 8+half of those needed to reach 8 in the preceding strata)
local ssize2=min(`ncat2',`ssizeC'+int((`ssizeC'-`ssize1')/2))
sample `ssize2', count, if strata==2 & stratmax==3


**** if 3 strata, the sample in strata two is equal to 24-sample in the first stratum
local ssize2b=min(`ncat2',`ssizeT'-`ssize1')
sample `ssize2b', count, if strata==2 & stratmax==2


*************************************
***SELECTION IN THIRD STRATUM**********
*************************************

*** NUMBER OF HOUSEHOLDS IN STRATUM 3
```

```
count if strata==3
local ncat3=r(N)

**** the sample in strata two is equal to 24-sample in the first two strata

local ssize3=min(`ncat3',`ssizeT'-`ssize1'-`ssize2')
sample `ssize3', count, if strata==3

sort numcat migstat id

sort migstat eanum strcnu hhdnum

**********************************************
***COMPUTE PROBABILITY OF SELECTION************
**********************************************
by migstat, sort: egen numsel=count(migstat)
gene prob2=numsel/numcat

**************************************
***CREATION OF HOUSEHOLD NUMBER********
**************************************

**************************************
***OUTFILE*****************************
**************************************
local fname=ea[1]
local fnum=eanum[1]

keep ea eanum strcnu address hhdnum headhhd partn retm nomig migstat comm prob2
save Accra_`fname'_select, replace
}

clear
use Accra_1_select
save Accra_1_cum, replace

foreach y of numlist 2/60 {

clear
local numf=`y'-1
use Accra_`numf'_cum

append using Accra_`y'_select
save Accra_`y'_cum, replace
}

clear
use Accra_60_cum
sort eanum strc hhd

gene id=_n
gene id_b=string(id) if id<10000
replace id_b="0"+string(id) if id<1000
replace id_b="00"+string(id) if id<100
replace id_b="000"+string(id) if id<10

drop id
saveold Accra_sample, replace
```

ANNEX 3: STATA SYNTAX FOR THE COMPUTATION OF WEIGHTS IN GHANA

```
*************************************************************
*** COMPUTATION OF HOUSEHOLD AND BIO SAMPLING WEIGHTS*********
********************** MAFE GHANA*****************************
*** *** BRUNO SCHOUMAKER _ NEW VERSION 21/09/2011 **************
*************************************************************

*** INCLUDES CORRECTION OF VARIABLES
*** WEIGHTS ARE CORRECTED FOR NON_RESPONSE
*** TRIMMING OF WEIGHTS SO THAT RATIO OF EXTREME WEIGHTS <100
*** WEIGHTS ARE ADJUSTED FOR POPULATION  SIZE

*** open household data file

cd "G:\COPIE PC BUREAU\Bruno\0.MAFE\Sampling\SAMPLING ACCRA_KUMASI\"
use "qm_household.dta", clear

replace n_menage=n_ques if n_menage==.
drop if n_menage==.

save "qm_household.dta", replace

**** Correction of DR number based on  sampling frame

replace num_dr=1 if num_dr==43 & n_men==19
replace num_dr=6 if num_dr==16 & n_men==125
replace num_dr=7 if num_dr==3 & n_men==155
replace num_dr=19 if num_dr==2 & n_men==440
replace num_dr=19 if num_dr==6 & n_men==452
replace num_dr=21 if num_dr==2 & n_men==487
replace num_dr=24 if num_dr==23 & n_men==560
replace num_dr=24 if num_dr==23 & n_men==565
replace num_dr=24 if num_dr==1 & n_men==569
replace num_dr=27 if num_dr==4 & n_men==648
replace num_dr=38 if num_dr==28 & n_men==889
replace num_dr=41 if num_dr==.b & n_men==980
replace num_dr=44 if num_dr==4 & n_men==1051
replace num_dr=49 if num_dr==21 & n_men==1172
replace num_dr=53 if num_dr==21 & n_men==1258
replace num_dr=60 if num_dr==3 & n_men==1440
replace num_dr=62 if num_dr==63 & n_men==1470
replace num_dr=72 if num_dr==73 & n_men==1723
replace num_dr=74 if num_dr==75 & n_men==1756
replace num_dr=74 if num_dr==75 & n_men==1768
replace num_dr=75 if num_dr==43 & n_men==1787
replace num_dr=75 if num_dr==42 & n_men==1788
replace num_dr=75 if num_dr==76 & n_men==1790

sort n_menage

save t_general, replace

use "SAMPLE_HOUSEHOLDS_ACCRA_KUMASI.DTA", clear
rename ide n_menage
destring, replace
sort n_men
```

```
gene migstat=0
replace migstat=1 if PARTN=="Y"
replace migstat=2 if RETU=="Y"

label var migstat "Strata - migration status"
label define migstat 0 "non migrant" 1 "Partners of migrant" 2 "Return migrants"
label val migstat migstat

save sample, replace

merge n_men using t_general

sort EA_NUM
cap drop _m
save t_general_m, replace

use "SAMPLE_EA_ACCRA_KUMASI.DTA", clear
destring, replace
rename EA_MAFE EA_NUM
sort EA_NUM

merge 1:m EA_NUM using t_general_m

gene proba_all=Selproba*Sampl

gene weight_hh1=1/proba_all

*** keep the strata

keep n_men EA_NUM weight_hh1 migsta Selproba Sampl proba_all

label var Selproba "P1.Probability of selection of PSU"
label var Sampl "P2. Probability of selection of household within PSU"
label var proba_all "PA. Probability of selection of household (P1*P2)"
label var weight_hh1 "HSW1. Househols sampling weight (1/PA)"

drop if EA_==.
save "weight_hh_MAFEGHANA.dta", replace

*****************************************************************
** Merging with sampling weights of households - before correction
*****************************************************************
use "qm_household.dta", clear

**** Correction of DR number - based on DR in sampling frame

replace num_dr=1 if num_dr==43 & n_men==19
replace num_dr=6 if num_dr==16 & n_men==125
replace num_dr=7 if num_dr==3 & n_men==155
replace num_dr=19 if num_dr==2 & n_men==440
replace num_dr=19 if num_dr==6 & n_men==452
replace num_dr=21 if num_dr==2 & n_men==487
replace num_dr=24 if num_dr==23 & n_men==560
replace num_dr=24 if num_dr==23 & n_men==565
replace num_dr=24 if num_dr==1 & n_men==569
replace num_dr=27 if num_dr==4 & n_men==648
replace num_dr=38 if num_dr==28 & n_men==889
replace num_dr=41 if num_dr==.b & n_men==980
replace num_dr=44 if num_dr==4 & n_men==1051
```

```
replace num_dr=49 if num_dr==21 & n_men==1172
replace num_dr=53 if num_dr==21 & n_men==1258
replace num_dr=60 if num_dr==3 & n_men==1440
replace num_dr=62 if num_dr==63 & n_men==1470
replace num_dr=72 if num_dr==73 & n_men==1723
replace num_dr=74 if num_dr==75 & n_men==1756
replace num_dr=74 if num_dr==75 & n_men==1768
replace num_dr=75 if num_dr==43 & n_men==1787
replace num_dr=75 if num_dr==42 & n_men==1788
replace num_dr=75 if num_dr==76 & n_men==1790

sort n_menage
save, replace

**** merge with weight file******
use "weight_hh_MAFEGHANA.dta", clear

sort n_men
merge n_menage using "qm_household.dta"

gene interv=1
replace interv=0 if ident==""

***********************************************
*computation of response rate by migstat and EA
***********************************************

*** COMPUTE EXPECTED AND REAL NUMBER OF HOUSEHOLDS IN EACH EA AND STRATA

gene eli_hh=1
gene int_hh=1 if interv==1

by EA migstat, sort : egen neli_hh=sum(eli_hh)
by EA migstat, sort : egen nint_hh=sum(int_hh)

*** compute participation rate

gene part_rate=nint_hh/neli_hh

label var part_rate "P3.Response rate by EA and Migration Status"

gene proba_final=proba_all*part_rate

label var proba_final "PF:P1*P2*P3"

gene weight_hh2=1/proba_final

label var weight_hh2 "HSW2. Household sampling weight (1/PF)- including non responses "

keep num_dr migstat n_men weight_hh1 weight_hh2 migsta Selproba Sampl proba_all part_rate

keep if num_dr!=.
keep if weight_hh2<.

sort n_men

**********************
*** Trimming weights
**********************
```

```
sort weight_hh2

local i = 1
while (weight_hh2[_N-`i'+1]/weight_hh2[`i']) > 100 {
local ratio=weight_hh2[_N-`i'+1]/weight_hh2[`i']
display `i' "  " `ratio'
local i = `i' + 1
}
local ratio=weight_hh2[_N-`i'+1]/weight_hh2[`i']
display "valeur finale " `i' "  " `ratio'

gene weight_hh3=weight_hh2

label var weight_hh3 "HSW3. Household sampling weight (1/PF)- including non responses - trimmed weights"

*replace by trimmed values
forvalues x = 1(1)`i' {
replace weight_hh3=weight_hh3[_N-`i'+1] if _n==(_N-`x'+1)
replace weight_hh3=weight_hh3[`i'] if _n==`x'
}

sort n_men
cap drop _m

save "weight_hh_MAFEGHANA_FULL.dta", replace

***** make sure the sum of weights = population size
use "G:\COPIE PC BUREAU\Bruno\0.MAFE\Sampling\SAMPLING ACCRA_KUMASI\qm_persons.dta", clear

replace n_menage=n_ques if n_menage==.
drop if n_menage==.

cap drop _m
sort n_menage

merge n_menage using "weight_hh_MAFEGHANA_FULL.dta"

total weight_hh3 if a4==1

** sum of weight= 1407591
** World urbanization prospects (2009) : Accra = 2269000; Kumasi=1773000 --> TOTAL=4 952 000

* Accra - 917443
total weight_hh3 if num_dr<61 & a4==1

* Kumasi - 490148
total weight_hh3 if num_dr>60 & a4==1

use "weight_hh_MAFEGHANA_FULL.dta", clear
cap drop _m
drop if n_menage==.

gene corr_fac=2269000/917443 if num_dr<61
replace corr_fac=1773000/490148 if num_dr>60

replace weight_hh3=weight_hh3*corr_fac
cap drop poidsmen_n
cap drop exp_fact_men
```

```
egen sum_w=sum(weight_hh3)
gene poidsmen_n=weight_hh3/sum_w*_N

rename weight_hh3 exp_fact_men

label var poidsmen_n "Ponderation normalise par menage"
label var exp_fact_men "inflating factor household"

keep n_men poidsmen_n exp_fact_men

save "weights_hhd_ghana.dta", replace

save "G:\WP5\Data\Ghana\weights\weights_hhd_ghana.dta", replace

*****************************************************************
**** PREPARATION OF FILE OF ELIGIBLE MEMBERS - HOUSEHOLD
*****************************************************************

*** OPEN HOUSEHOLD FILE
use "G:\COPIE PC BUREAU\Bruno\0.MAFE\Sampling\SAMPLING ACCRA_KUMASI\qm_persons.dta", clear
replace n_menage=n_ques if n_menage==.
drop if n_menage==.

cap drop poidsmen

cap drop exp_fac
sort n_men

*** MERGE WITH WEIGHT FILE
cap drop _m
merge n_men using "weights_hhd_ghana.dta"

*************************************************************************
**** corrections of household file identified in probabilistic matching
*************************************************************************

**** Correction of gender - based on gender in bio questionnaire
replace a1=2 if n_menage==9 & n_indiv==1
replace a1=1 if n_menage==28 & n_indiv==1
replace a1=2 if n_menage==1811 & n_indiv==1
replace a1=2 if n_menage==1837 & n_indiv==1
replace a1=2 if n_menage==48 & n_indiv==3
replace a1=2 if n_menage==145 & n_indiv==3
replace a1=2 if n_menage==210 & n_indiv==1
replace a1=2 if n_menage==255 & n_indiv==1
replace a1=2 if n_menage==346 & n_indiv==1
replace a1=2 if n_menage==400 & n_indiv==1
replace a1=2 if n_menage==515 & n_indiv==1
replace a1=2 if n_menage==538 & n_indiv==1
replace a1=2 if n_menage==739 & n_indiv==2
replace a1=2 if n_menage==879 & n_indiv==1
replace a1=2 if n_menage==978 & n_indiv==1
replace a1=2 if n_menage==538 & n_indiv==2
replace a1=2 if n_menage==1074 & n_indiv==1
replace a1=2 if n_menage==1082 & n_indiv==2
replace a1=2 if n_menage==1085 & n_indiv==1
replace a1=2 if n_menage==1172 & n_indiv==1
replace a1=2 if n_menage==1205 & n_indiv==1
replace a1=2 if n_menage==1246 & n_indiv==1
replace a1=2 if n_menage==1366 & n_indiv==1
```

```
replace a1=2 if n_menage==1367 & n_indiv==1
replace a1=2 if n_menage==1431 & n_indiv==6
replace a1=2 if n_menage==1527 & n_indiv==1
replace a1=2 if n_menage==1637 & n_indiv==2
replace a1=1 if n_menage==1423 & n_indiv==1
replace a1=2 if n_menage==1601 & n_indiv==1
replace a1=2 if n_menage==1182 & n_indiv==2
replace a1=1 if n_menage==516 & n_indiv==1
replace a1=1 if n_menage==949 & n_indiv==1
replace a1=1 if n_menage==1639 & n_indiv==1

**** Correction of age - based on gender in bio questionnaire

replace q3age=42 if n_menage==176 & n_indiv==7
replace q3age=38 if n_menage==194 & n_indiv==1
replace q3age=24 if n_menage==1020 & n_indiv==6
replace q3age=38 if n_menage==578 & n_indiv==1

**** Correction of DR - based on sampling frame

replace num_dr=1 if num_dr==43 & n_men==19
replace num_dr=6 if num_dr==16 & n_men==125
replace num_dr=7 if num_dr==3 & n_men==155
replace num_dr=19 if num_dr==2 & n_men==440
replace num_dr=19 if num_dr==6 & n_men==452
replace num_dr=21 if num_dr==2 & n_men==487
replace num_dr=24 if num_dr==23 & n_men==560
replace num_dr=24 if num_dr==23 & n_men==565
replace num_dr=24 if num_dr==1 & n_men==569
replace num_dr=27 if num_dr==4 & n_men==648
replace num_dr=38 if num_dr==28 & n_men==889
replace num_dr=41 if num_dr==.b & n_men==980
replace num_dr=44 if num_dr==4 & n_men==1051
replace num_dr=49 if num_dr==21 & n_men==1172
replace num_dr=53 if num_dr==21 & n_men==1258
replace num_dr=60 if num_dr==3 & n_men==1440
replace num_dr=62 if num_dr==63 & n_men==1470
replace num_dr=72 if num_dr==73 & n_men==1723
replace num_dr=74 if num_dr==75 & n_men==1756
replace num_dr=74 if num_dr==75 & n_men==1768
replace num_dr=75 if num_dr==43 & n_men==1787
replace num_dr=75 if num_dr==42 & n_men==1788
replace num_dr=75 if num_dr==76 & n_men==1790


**** replace age in hhd survey by age in bio survey if difference <=10 ans

replace q3age=65 if n_menage==29 & n_indiv==1
replace q3age=59 if n_menage==29 & n_indiv==2
replace q3age=31 if n_menage==66 & n_indiv==1
replace q3age=41 if n_menage==70 & n_indiv==4
replace q3age=60 if n_menage==84 & n_indiv==5
replace q3age=51 if n_menage==138 & n_indiv==2
replace q3age=47 if n_menage==139 & n_indiv==1
replace q3age=30 if n_menage==154 & n_indiv==2
replace q3age=46 if n_menage==155 & n_indiv==1
replace q3age=43 if n_menage==155 & n_indiv==2
replace q3age=60 if n_menage==163 & n_indiv==1
replace q3age=60 if n_menage==169 & n_indiv==1
replace q3age=30 if n_menage==173 & n_indiv==1
```

```
replace q3age=50 if n_menage==208 & n_indiv==2
replace q3age=69 if n_menage==443 & n_indiv==1
replace q3age=49 if n_menage==312 & n_indiv==2
replace q3age=39 if n_menage==333 & n_indiv==1
replace q3age=65 if n_menage==980 & n_indiv==2
replace q3age=28 if n_menage==348 & n_indiv==5
replace q3age=28 if n_menage==340 & n_indiv==2
replace q3age=42 if n_menage==367 & n_indiv==1
replace q3age=29 if n_menage==380 & n_indiv==2
replace q3age=34 if n_menage==388 & n_indiv==1
replace q3age=53 if n_menage==401 & n_indiv==1
replace q3age=31 if n_menage==403 & n_indiv==3
replace q3age=62 if n_menage==459 & n_indiv==1
replace q3age=26 if n_menage==474 & n_indiv==2
replace q3age=30 if n_menage==516 & n_indiv==1
replace q3age=70 if n_menage==566 & n_indiv==1
replace q3age=51 if n_menage==636 & n_indiv==2
replace q3age=62 if n_menage==657 & n_indiv==1
replace q3age=29 if n_menage==708 & n_indiv==1
replace q3age=35 if n_menage==779 & n_indiv==1
replace q3age=51 if n_menage==845 & n_indiv==2
replace q3age=41 if n_menage==850 & n_indiv==1
replace q3age=47 if n_menage==865 & n_indiv==1
replace q3age=67 if n_menage==889 & n_indiv==2
replace q3age=40 if n_menage==889 & n_indiv==3
replace q3age=40 if n_menage==914 & n_indiv==1
replace q3age=32 if n_menage==949 & n_indiv==1
replace q3age=56 if n_menage==974 & n_indiv==1
replace q3age=54 if n_menage==974 & n_indiv==2
replace q3age=36 if n_menage==1109 & n_indiv==5
replace q3age=49 if n_menage==1010 & n_indiv==1
replace q3age=49 if n_menage==1091 & n_indiv==2
replace q3age=30 if n_menage==1048 & n_indiv==1
replace q3age=54 if n_menage==1054 & n_indiv==1
replace q3age=44 if n_menage==1054 & n_indiv==2
replace q3age=43 if n_menage==1092 & n_indiv==1
replace q3age=29 if n_menage==1140 & n_indiv==2
replace q3age=61 if n_menage==1172 & n_indiv==1
replace q3age=53 if n_menage==1174 & n_indiv==1
replace q3age=31 if n_menage==1180 & n_indiv==1
replace q3age=44 if n_menage==1182 & n_indiv==2
replace q3age=55 if n_menage==1219 & n_indiv==1
replace q3age=53 if n_menage==1237 & n_indiv==1
replace q3age=59 if n_menage==1258 & n_indiv==1
replace q3age=64 if n_menage==1470 & n_indiv==1
replace q3age=34 if n_menage==1622 & n_indiv==2
replace q3age=57 if n_menage==1494 & n_indiv==1
replace q3age=33 if n_menage==1526 & n_indiv==2
replace q3age=51 if n_menage==1623 & n_indiv==1
replace q3age=30 if n_menage==1655 & n_indiv==1
replace q3age=49 if n_menage==1662 & n_indiv==1
replace q3age=60 if n_menage==1692 & n_indiv==1
replace q3age=28 if n_menage==1723 & n_indiv==1
replace q3age=36 if n_menage==1748 & n_indiv==2
replace q3age=27 if n_menage==1876 & n_indiv==3
replace q3age=61 if n_menage==1768 & n_indiv==1
replace q3age=54 if n_menage==1768 & n_indiv==2
replace q3age=39 if n_menage==1617 & n_indiv==1
replace q3age=39 if n_menage==1844 & n_indiv==1
replace q3age=30 if n_menage==1842 & n_indiv==2
```

```
replace q3age=35 if n_menage==1899 & n_indiv==4
replace q3age=30 if n_menage==1916 & n_indiv==1


**** ELIGIBLES PERSONS IN HOUSEHOLDS ***


*********************************************
*** eligible return migrants *****************
*********************************************

tab a13c q4_return, m

gene migret=0

* Return migrant
replace migret=1 if q4_return==1
** aged betwee 25 and 75
replace migret=0 if q3age<25 | q3age >75
** born in Ghana
replace migret=0 if a14pays!=99329
** curently living in the household
replace migret=0 if a4!=1


*********************************
** partners/spouses of migrant **
*********************************

gene partmig=0

* partner
replace partmig=1 if q4_cjt==1
** aged between 25 and 75
replace partmig=0 if q3age<25 | q3age >75
** born in Ghana
replace partmig=0 if a14pays!=99329
** curently living in the household
replace partmig=0 if a4!=1

*** priority to return migrant (si partenaire et migrant de retour, considere comme migrant de
retour)
replace partmig=0 if migret==1

*****************************
** other eligible members
*****************************

gene otherel=0

* all people except partners and return migrants
replace otherel=1 if partmig==0 & migret==0
** aged between 25 and 75
replace otherel=0 if q3age<25 | q3age >75
** born in Ghana
replace otherel=0 if a14pays!=99329
** curently living in the household
replace otherel=0 if a4!=1
```

```
*** check that sum of strata=1
gene sumstrate=migret+partmig+otherel


*******************************************************************************
*** computation of number of eligible members in each household for each stratum
*******************************************************************************

by n_men, sort : egen nmigret=sum(migret)
by n_men, sort : egen npartmig=sum(partmig)
by n_men, sort : egen notherel=sum(otherel)


list n_men n_indi migret partmig otherel q3age a14pays a4

******************************************************
** creation of two variables for probabilistic matching
******************************************************

gene sex=a1
gene age=q3age

keep ident n_indiv n_menage age migret partmig otherel sex num_dr npart nmig nothe poidsmen
exp* q3age a14pays a4

** id for probabilistic matching
sort n_menage n_indiv
gene idM=_n

save hh_ghana_elig.dta, replace

*******************************
* Save file of eligible persons
*******************************

by n_men, sort : gene eli_other=1 if notherel>0

by n_men, sort : gene first=1 if _n==1

tab eli_other if first==1


****************************************
******** OPEN BIO DATA FILE *******
****************************************
*** ouverture base bio

use "G:\COPIE PC BUREAU\Bruno\0.MAFE\Sampling\SAMPLING
ACCRA_KUMASI\GH_qb_general_110719.dta", clear
replace id_coun=paysenq if id_coun==""
save, replace
keep if id_coun=="G"


replace num_dr=nodr if num_dr==.

gene migret=0
replace migret=1 if q600m>0

**** corrections after probabilistic matching
```

```
replace q1=2 if n_menage==944 & n_indiv==2

*** Correction of individual numbers based on hhd questionnaire

gene n_indiv_old=n_indiv

replace n_indiv=2 if n_menage==1036 & n_indiv==3
replace n_indiv=1 if n_menage==1195 & n_indiv==224
replace n_indiv=1 if n_menage==1602 & n_indiv==0
replace n_indiv=1 if n_menage==1753 & n_indiv==0
replace n_indiv=2 if n_menage==1754 & n_indiv==0
replace n_indiv=1 if n_menage==1755 & n_indiv==0
replace n_indiv=2 if n_menage==1757 & n_indiv==0
replace n_indiv=1 if n_menage==1758 & n_indiv==0
replace n_indiv=1 if n_menage==1759 & n_indiv==0
replace n_indiv=2 if n_menage==1762 & n_indiv==0
replace n_indiv=1 if n_menage==1763 & n_indiv==0
replace n_indiv=1 if n_menage==1764 & n_indiv==0
replace n_indiv=1 if n_menage==1766 & n_indiv==0
replace n_indiv=2 if n_menage==1770 & n_indiv==0
replace n_indiv=2 if n_menage==1772 & n_indiv==0
replace n_indiv=1 if n_menage==1773 & n_indiv==0
replace n_indiv=1 if n_menage==1774 & n_indiv==0
replace n_indiv=1 if n_menage==1775 & n_indiv==0
replace n_indiv=1 if n_menage==1776 & n_indiv==0
replace n_indiv=1 if n_menage==1810 & n_indiv==0
replace n_indiv=2 if n_menage==1823 & n_indiv==3


replace n_indiv=2 if n_menage==14 & n_indiv==1
replace n_indiv=3 if n_menage==95 & n_indiv==2
replace n_indiv=1 if n_menage==175 & n_indiv==17
replace n_indiv=2 if n_menage==221 & n_indiv==1
replace n_indiv=1 if n_menage==229 & n_indiv==10

replace n_indiv=2 if n_menage==790 & n_indiv==1
replace n_indiv=2 if n_menage==947 & n_indiv==1
replace n_indiv=1 if n_menage==221 & n_indiv==2
replace n_indiv=2 if n_menage==222 & n_indiv==1
replace n_indiv=2 if n_menage==1093 & n_indiv==1
replace n_indiv=1 if n_menage==1109 & n_indiv==2
replace n_indiv=2 if n_menage==1513 & n_indiv==3
replace n_indiv=2 if n_menage==1637 & n_indiv==1
replace n_indiv=6 if n_menage==1736 & n_indiv==5
replace n_indiv=1 if n_menage==333 & n_indiv==14


sort n_menage n_indiv
gene idB=_n

gene age=2009-q1a
gene sex=q1
save bio_ghana.dta, replace


*****************************************
**** Probabilistic matching with household
*****************************************
```

```
use bio_ghana.dta, clear
cap drop _m

** PROBABILISTIC MATCHING BASED ON 6 VARIABLES

reclink num_dr n_menage n_indiv sex age migret using hh_ghana_elig.dta, gen(myscore) idm(idB)
idu(idM) wmatch(10 15 10 5 5 5) wnomatch(10 20 5 1 10 5)

** identify households with mimatch
by n_menage, sort: egen mismatch=min(myscore)
sort n_menage n_indiv

list n_menage Un_menage n_indiv Un_indiv sex Usex age Uage migret Umigret myscore if
myscore<0.99, c
list n_menage n_indiv Umigret nmigret partm npartm otherel notherel


**********
*** computation of number of eligible members in each household
**********

by n_men, sort : egen nmigret_r=sum(Umigret)
by n_men, sort : egen npartmig_r=sum(partmig)
by n_men, sort : egen notherel_r=sum(otherel)


list n_menage n_indiv nmigret_r nmigret npartmig_r npartmig notherel_r notherel

*** Computation of sampling rate in each stratum

gene sam_migret=nmigret_r/nmigret
replace sam_migret=0 if sam_migret==.
gene sam_partmig=npartmig_r/npartmig
replace sam_partmig=0 if sam_partmig==.
gene sam_notherel=notherel_r/notherel
replace sam_notherel=0 if sam_notherel==.

list n_menage n_indiv sam*

*** Attribute sampling rate to individual accoridng to stratum

gene sumstrate=Umigret+partmig+otherel

gene sam_rate=Umigret*sam_migret+partmig*sam_partmig+otherel*sam_notherel

list n_menage n_indiv sam_rate Umigret sam_migret partmig sam_partmig otherel sam_notherel


*** check if all individuals are eligible

gene weight_bio_1=1/sam_rate
replace weight_bio_1=0 if sam_rate==0

gene final_w_bio=weight_bio*exp_fact_men

*** dropper 8 individuals not eligible

replace a14p=99329 if a14p==.
list ident if a14p!=99329
drop if a14p!=99329
```

```
list ident if q1a==1985
drop if q1a==1985

*** random imputation for others (20 missinga and weight=0)

set seed 100
gene rnum=int(runiform()*_N)
replace final_w_bio=final_w_bio[rnum] if final_w_bio==. | final_w_bio==0

*****

list n_menage n_indiv age Umigret partmig otherel nmigret_r nmigret npartmig_r npartmig
notherel_r notherel if sam_rate==0, c

tab1 q3age a14pays a4 if weight_bio!=0


*--------- check total
total final_w_bio

matrix c=e(b)'
svmat c, name(total_bio)
scalar totalbio=total_bio[1]
di totalbio

**********
*** Trimming weights
**********

sort final_w_bio

local i = 1
while (final_w_bio[_N-`i'+1]/final_w_bio[`i']) > 100 {
local ratio=final_w_bio[_N-`i'+1]/final_w_bio[`i']
display `i' "  " `ratio'
local i = `i' + 1
}
local ratio=final_w_bio[_N-`i'+1]/final_w_bio[`i']
display "valeur finale " `i' "  " `ratio'

gene bio_weight_3=final_w_bio

label var bio_weight_3 "BW3. Biographic sampling weight - including non responses - trimmed
weights"


*replace by trimmed values
forvalues x = 1(1)`i' {
replace bio_weight_3=bio_weight_3[_N-`i'+1] if _n==(_N-`x'+1)
replace bio_weight_3=bio_weight_3[`i'] if _n==`x'
}

keep n_menage n_indiv_old bio_weight_3

sort n_menage n_indiv_old n_indiv

*--------- check total
total bio_weight_3
```

```
matrix c=e(b)'
svmat c, name(total_bio2)
scalar totalbio2=total_bio2[1]
di totalbio2

*** adjust expansion weights for trimming
replace bio_weight_3=bio_weight_3*totalbio/totalbio2
total bio_weight_3
save "weight_bio_MAFEGHANA_FULL.dta", replace


**** merge with bio file
use "G:\COPIE PC BUREAU\Bruno\0.MAFE\Sampling\SAMPLING
ACCRA_KUMASI\GH_qb_general_110719.dta", clear
*use "Y:\Mafe\MAFE-FP7\WP5\DATA\Bases MAFE Bio Ghana\qb_general.dta", clear
cap drop _m
keep if id_coun=="G"
gene n_indiv_old=n_indiv
sort n_menage n_indiv_old

merge n_menage n_indiv_old using  "weight_bio_MAFEGHANA_FULL.dta"

drop n_indiv_old
rename bio_weight_3 wei_gh_bio

keep ident wei_gh_bio
sort iden
save "weight_bio_MAFEGHANA_FULL_small.dta", replace


**** ********** WEIGHTS IN EUROPE **********

use "G:\COPIE PC BUREAU\Bruno\0.MAFE\Sampling\SAMPLING
ACCRA_KUMASI\GH_qb_general_110719.dta", clear
sort ident
cap drop _m
merge iden using "weight_bio_MAFEGHANA_FULL_small.dta"

gene age=2009-q1a

recode age (0/24=0) (25/34=1) (35/44=2) (45/54=3) (55/110=4), gen(ageg)
recode age (0/24=0) (25/34=1) (35/49=2) (50/64=3) (65/75=4), gen(ageg_UK)


replace wei_gh_bio=27 if ageg==1 & q1==1 & id_c=="N"
replace wei_gh_bio=28 if ageg==2 & q1==1 & id_c=="N"
replace wei_gh_bio=66 if ageg==3 & q1==1 & id_c=="N"
replace wei_gh_bio=45 if ageg==4 & q1==1 & id_c=="N"

replace wei_gh_bio=31 if ageg==1 & q1==2 & id_c=="N"
replace wei_gh_bio=69 if ageg==2 & q1==2 & id_c=="N"
replace wei_gh_bio=63 if ageg==3 & q1==2 & id_c=="N"
replace wei_gh_bio=18 if ageg==4 & q1==2 & id_c=="N"

*** 8 persons born in 1985 or later in Netherlands - included in 25 yrs

replace wei_gh_bio=27 if ageg==0 & q1==1 & id_c=="N"
replace wei_gh_bio=31 if ageg==0 & q1==2 & id_c=="N"
```

```
replace wei_gh_bio=302 if ageg_UK==1 & q1==1 & id_c=="U"
replace wei_gh_bio=745 if ageg_UK==2 & q1==1 & id_c=="U"
replace wei_gh_bio=550 if ageg_UK==3 & q1==1 & id_c=="U"
replace wei_gh_bio=434 if ageg_UK==4 & q1==1 & id_c=="U"

replace wei_gh_bio=510 if ageg_UK==1 & q1==2 & id_c=="U"
replace wei_gh_bio=456 if ageg_UK==2 & q1==2 & id_c=="U"
replace wei_gh_bio=425 if ageg_UK==3 & q1==2 & id_c=="U"
replace wei_gh_bio=687 if ageg_UK==4 & q1==2 & id_c=="U"


keep ident wei_gh_bio id_c
sort ident
save "weight_bio_MAFEGHANA_FULL_small_2.dta", replace


*** Normalized weights per country / poid_norm_ctry
*** sum of weights = sample size

by id_c, sort : egen tot_pays=sum(wei_gh_bio)
by id_c, sort : egen sam_pays=count(wei_gh_bio)
gene poid_norm_ctry=wei_gh_bio/tot_pays*sam_pays


*** Normalized weights Europe / poid_norm_eur
*** sum of weights = sample size in Europe

egen tot_euro=sum(wei_gh_bio) if id_c=="U"|id_c=="N"
egen sam_euro=count(wei_gh_bio) if id_c=="U"|id_c=="N"
gene poid_norm_eur=wei_gh_bio/tot_euro*sam_euro


*** Normalized weights All / poid_norm_all
*** sum of weights = sample size in Europe+Africa

egen tot_all=sum(wei_gh_bio)
egen sam_all=count(wei_gh_bio)
gene poid_norm_all=wei_gh_bio/tot_all*sam_all

keep ident wei_gh_bio id_c poid_norm_ctry poid_norm_eur poid_norm_all
sort ident
save "weight_bio_MAFEGHANA.dta", replace

** remove doublons
by ident, sort: keep if _n==1

save "G:\WP5\Data\Ghana\weights\weight_bio_MAFEGHANA.dta", replace
```